LES CAHIERS DE L'ANALYSE DES DONNÉES

J. P. BENZÉCRI

Histoire et préhistoire de l'analyse des données. Partie III Era piscatoria

Les cahiers de l'analyse des données, tome 1, nº 3 (1976), p. 221-241

<http://www.numdam.org/item?id=CAD_1976__1_3_221_0>

© Les cahiers de l'analyse des données, Dunod, 1976, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme Numérisation de documents anciens mathématiques

http://www.numdam.org/

Les Cahiers de l'Analyse des Données
Vol. I - 1976 - n° 3 - p. 221-241

HISTOIRE ET PRÉHISTOIRE DE L'ANALYSE DES DONNÉES

Partie III - Era piscatoria

par J. P. Benzécri (1)

2.3. Era Piscatoria :

Quand en 1936 Karl Pearson quitta ce monde, G.U. Yule, (1871-1951) put s'exclamer : "I feel as though the Karlovingian era has come to an end, and the Piscatorial era which succeeds it is one in which I can play no part". En effet, des mains de Karl, le sceptre de la statistique était passé au pêcheur c'est-à-dire à R.A. Fisher (1890-1962). Comme K. Pearson, Sir Ronald Fisher fut un penseur original, un docteur de la science méprisant la pénombre des compromis. En presque tout, autant qu'il le put, Fisher s'opposa à Pearson; mais maintenant que le silence s'est fait sur le champ de bataille, on peut affirmer sereinement que celui-là fut le continuateur de celui-ci.

Du point de vue de l'analyse des données, l'oeuvre de R.A. Fisher est précieuse en ce qu'elle a accoutumé les statisticiens à la géométrie multidimensionnelle. Mais les règles de Fisher pour l'induction (§ 2.3.6) fondées sur une expression exacte des lois probabilistes, l'ont conduit selon nous à trop attendre du modèle normal (*). C'est pourquoi l'analyse des données assistée de l'ordinateur, qui traite telles quelles des distributions empiriques - nuages de points - s'est dans ses débuts heurtée souvent avec violence à la méthode issue de Fisher. L'avenir dira si cette méthode, renouvelée, peut apporter à l'analyse des données des épreuves de validité pratiques et non illusoires (§ 3.8.5).

(*) *Fisher lui-même dans sa communication de 1924 au Congrès de Toronto (Proceedings of the International Congress of Math, Toronto; T. 2 pp. 805-813; et Contributions) témoigne qu'il n'accepte pas en aveugle le postulat de normalité quand il déplore que "l'étudiant n'est pas d'ordinaire averti... que la distribution des erreurs expérimentales n'a qu'un lointain rapport avec la loi normale; laquelle ne vaut dans la pratique que parce qu'on l'applique non à des observations isolées, mais à des moyennes ou d'autres statistiques calculées sur plusieurs observations". Mais nous ne croyons pas que le théorème central limite (cf § 1.4.3) suffise à légitimer le modèle normal multidimensionnel dont la forme des distributions empiriques s'écarte même qualitativement (on l'a dit au § 2.2.5 en étudiant les corrélations partielles).*

(**) *Suite des articles parus sous le même titre dans les cahiers n° 1 & 2, vol. I, pp. 9-37 & 101-120.*

(1) *Professeur ; Laboratoire de Statistique - Université Pierre et Marie Curie - Paris.*

Après Fisher, d'éminents statisticiens -citons E.S. Pearson, J. Neyman, A. Wald, M.G. Kendall - ont poursuivi son oeuvre parfois dans un autre esprit; mais nous n'en dirons presque rien dans le présent exposé qui vise à éclairer l'histoire de l'analyse des données. L'oeuvre si considérable de Fisher (cinq livres; un recueil de tables, établi avec F. Yates; des centaines d'articles de statistique et de génétique rassemblés dans les *Collected papers of R.A. Fisher* édités par l'Université d'Adélaïde; et dont une anthologie statistique, publiée en 1950 par J. Wiley et Chapman & Hall sous le titre de *Contributions to mathematical statistics* et citée ici *Contributions*, est dans bien des bibliothèques) ne pouvait elle-même être résumée comme elle le mérite. Nous nous bornerons à en énumérer les principales parties en précisant leur rapport avec notre objet : mais comme les spécialistes de l'analyse des données auxquels nous destinons cet exposé sont souvent peu avertis des problèmes traités par Fisher, nous croyons nécessaire de présenter ceux-ci avec concision en usant du langage algébrique et géométrique aujourd'hui très répandu.

2.3.1. Mendel et Galton : Comme K. Pearson, R.A. Fisher fut toute sa vie passionné par la génétique (et aussi par la doctrine sociale souvent inquiétante que sous le nom d'eugénisme on prétend déduire logiquement de celle-ci). Plus heureux en cela que Pearson, Fisher vint à l'heure où (notamment grâce aux expériences de Morgan sur la drosophile) les lois de Mendel apparaissaient déjà une acquisition confirmée de la science; et il eut le grand mérite de conjuguer rationnellement en une seule doctrine la sélection naturelle, les mutations, la transmission mendélienne des caractères et les formules de variance et de covariance de la biométrie. Dès 1918 R.A. Fisher étudie "The correlation between relatives on the supposition of Mendelian inheritence" (*Trans. Roy. Soc. Edimb.* T. 52; pp. 399-443) : en assouplissant (comme l'avait déjà suggéré G.U. Yule) la règle mendélienne de dominance (selon laquelle Dr a exactement la même apparence que DD) en supposant (Pearson et Weldon l'avaient déjà fait : cf § 2.2.4) qu'un caractère continu visible (e.g. la taille) dépend de plusieurs caractères génétiques mendéliens (locus), Fisher trouve pour les corrélations entre parents et enfants, entre frères, etc..., des valeurs compatibles avec la *Law of Ancestral Heredity*. Cette loi démontre qu'une partie de la diversité qu'on observe entre les individus d'une même génération est héritée de leurs parents : en ce sens, elle démontre l'existence de ce qu'on peut d'un terme vague appeler *capital génétique* et affirme que ce capital varie d'individu à individu. Cependant sous la théorie de l'enfant moyenne de ses parents ce capital tendrait à s'uniformiser par le jeu égalisateur des mariages. Fisher montre clairement que seule l'hérédité mendélienne (nous dirons quantifiée; par transmission d'unités discontinues d'information) permet que se maintienne entre individus la diversité du capital génétique, sans qu'il soit besoin d'admettre l'hérédité des caractères acquis, ni un taux de mutation bien supérieur à celui observé (cf *The biometrical study of heredity*, in *Eugenics Review*, T.16, pp. 189-210; 1924; notamment : pp. 203-204).

Dans la suite, les recherches génétiques de Fisher (qui occupèrent une noble part de son génie : Fisher fut à Cambridge professeur de génétique; tandis qu'il n'occupa jamais de chaire de statistique (*)) s'écartèrent de plus en plus des problèmes posés par l'Ecole biométrique : aussi n'en dirons-nous rien. Mais, c'est encore un problème de biométrie, la comparaison des races (ou des espèces d'un même genre, etc.). qui fut l'occasion des travaux de Fisher sur la discrimination; travaux dont nous parlerons au § 2.3.5.

(*) *Haldane (Biometrika T. 44, 1957 et Studies p. 443) écrit : "It is remarkable that the greatest statistician in the world never held a chair in statistics".*

2.3.2. *La statistique des petits échantillons* : il importe souvent dans la pratique de décider si une variable aléatoire x (par exemple l'amélioration apportée par un traitement) diffère significativement de zéro. Si l'on accepte le modèle normal on considérera donc l'hypothèse : x est distribué normalement avec pour moyenne zéro (et pour variance une grandeur inconnue σ). Une première idée se présente à l'esprit pour éprouver cette hypothèse : soit $x^I=\{x^1,\ldots,x^i,\ldots,x^n\}\in R^n$, un échantillon de n valeurs de la grandeur aléatoire x (e.g. : les améliorations - positives ou négatives - enregistrées dans n cas); on calcule la moyenne empirique $\bar{x} = (1/n)\ \Sigma\ \{x^i | i = 1,\ \ldots,\ n\}$; cette moyenne est distribuée autour de zéro avec pour variance (σ^2/n); mais σ^2 est inconnu ! on l'estime d'après la variance de l'échantillon : $(\sigma')^2 = (1/n)\ \Sigma\ \{(x^i - \bar{x})^2 | i = 1,\ \ldots,\ n\}$. On décide alors de conserver l'hypothèse de moyenne nulle si (e.g.) $\bar{x}/(\sigma' n^{-1/2})$ (qu'on considère selon l'hypothèse comme une variable normale centrée de variance 1) est inférieur à 3 en valeur absolue; sinon on concluera que $\bar{x}$ diffère significativement de zéro. Voilà à peu près comment opérait l'école biométrique au début du XX° siècle; de tels raisonnements avaient le grand mérite de suggérer une réponse à la question statistique fondamentale : cette structure que nous avons remarquée en dépouillant les données, est-elle réelle ? ou n'est-elle qu'une apparence produite par les fluctuations d'échantillonnage ? Cependant ces raisonnements n'étaient pas d'une rigueur mathématique irréprochable; et dans le cas de petits échantillons (n petit) conduisaient à des décisions peu en accord avec la réalité. Dès 1908 le chimiste Gosset, ingénieur dans une brasserie, publiait sous le nom modeste de Student (l'étudiant) un mémoire "The probable error of a mean" (*Biometrika* T. 6) qui est à l'origine de la détermination exacte des lois probabilistes de nombreuses grandeurs (ou : statistiques) calculées même sur de petits échantillons. Grandement intéressé par certaines données (celles de la brasserie) très patient dans la contemplation des tableaux de nombres, et leur confrontation aux formules les plus diverses (Pearson en avait proposé un précieux arsenal) Gosset nous rappelle Galton, découvrant la corrélation et la régression (§2.2.2). Quelque attachante que soit la figure de Gosset (cf L. Mc Mullen & E.S. Pearson, in *Biometrika* T. 30, 1939; reproduit dans *Studies*) il est plus conforme à notre propos de donner de la formule de "Student" une démonstration inspirée de Fisher.

Pour Fisher, sous l'hypothèse que x est distribué normalement avec moyenne nulle et variance inconnue σ^2, l'échantillon $\{x^1,\ \ldots,\ x^n\}$ est un point x^I de R^n (*) distribué suivant une loi normale sphérique centrée à l'origine (et pour laquelle chaque coordonnées a variance σ^2). La moyenne $\bar{x}$ (plus précisément son produit par $n^{1/2}$) est la coordonnée de x^I sur un nouvel axe Δ que nous appellerons diagonal car il est le lieu des points dont toutes les coordonnées sont égales : la quantité $n(\sigma')^2 = \Sigma\{(x^i - \bar{x})^2\}$, n'est autre que le carré de la distance du point x^I à cet axe Δ; et le vecteur $x^H = \{(x^i - \bar{x}) | i = 1,\ \ldots;\ n\}$, est la projection du vecteur x^I sur l'hyperplan H (sous-espace linéaire de R^n, de dimension n - 1) perpendiculaire à l'axe diagonal Δ. Dans H, $\{(x^i - \bar{x})\}$

(*) *Nous avons déjà fait usage de cette représentation si féconde, due croyons-nous à R. Fisher, (Biometrika, T. 10 pp. 507-521) pour exposer le principe de la méthode des moindres carrés (§ 1.5.2); et en donnerons au § 2.3.4 une description plus complète.*

est normal sphérique, à n - 1 dimensions, de variance σ^2; donc la meilleure estimation $\tilde{\sigma}^2$ de σ^2 n'est pas $\Sigma\{(x^i - \bar{x})^2\}/n$, mais $\Sigma\{x^i - \bar{x})^2\}/(n - 1)$. La loi du rapport $t = n^{1/2}\,\bar{x}/\tilde{\sigma} = \bar{x}((n(n - 1)/\Sigma\{(x^i - \bar{x})^2\})^{1/2}$ (Student considère le rapport t parce que à la limite, pour n infini, $\tilde{\sigma} \# \sigma$ et $n^{1/2}\bar{x}$ est normal centré de variance σ; donc t est normal centré de variance 1), s'obtient sans référence directe à la loi normale, par la seule considération de la symétrie sphérique de la loi du vecteur x^I : la loi de :

$$n^{1/2}\,\bar{x}/\Sigma\{(x^i - \bar{x})^2\})^{1/2} = (n - 1)^{-1/2}\, t;$$

n'est autre que celle du rapport $|x^\Delta|/|x^H|$ des composantes (dans la direction d'un axe Δ et de l'hyperplan H qui lui est perpendiculaire) d'un vecteur x^I dont l'extrémité est uniformément distribuée sur une sphère Σ de R^n; c'est ce que symbolise la figure 2-4.

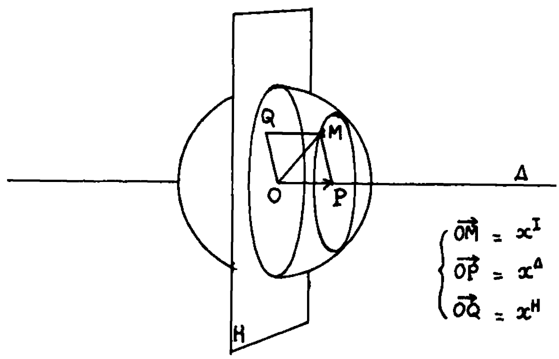

Figure 2-4 : Décomposition, dans deux directions supplémentaires Δ et H, d'un vecteur $\vec{OM}$ dont l'extrémité est uniformément distribuée sur une hypersphère.

L'exemple qui précède est typique à plus d'un titre. D'abord, il est le plus simple d'une longue série de résultats sur la distribution exacte de statistiques issues de petits échantillons : ainsi, dès 1915 (*Biometrika*, T. 10); Fisher (par un raisonnement géométrique analogue mais plus complexe) détermine la distribution exacte du coefficient de corrélation empirique r entre x et y, calculé sur un échantillon $\{(x^1, y^1),\ldots, (x^n, y^n)\}$ d'effectif n, issu d'une loi normale bidimensionnelle pour laquelle la corrélation est ρ. Ensuite, l'idée de considérer un ensemble de nombres (données mesurées : e.g. les $\{x^i\}$; ou les 2n nombres x^i et y^i pris ensemble) comme les coordonnées d'un point (de R^n ou R^{2n}) dont la distribution est normale sphérique; puis de faire des changements d'axes dans l'espace, est à la base de l'analyse de la variance telle que l'a conçue Fisher (cf 2.3.4); et une représentation analogue permet de regarder l'estimation du maximum de vraisemblance (§ 2.3.3) comme une sorte de projection orthogonale (fig. 2-6).

Nous admirons les représentations géométriques dont use Fisher. Nous reparlerons de l'analyse de la variance. Quant à la statistique des petits échantillons, il ne fait pas de doute qu'une distribution exacte soit préférable à une approximation en partie incorrecte ! toutes deux étant fondées sur une hypothèse de normalité ne devraient toutefois avoir dans la pratique qu'une valeur indicative. De ce point de vue les petits échantillons sont des échantillons insuffisants; et si n est assez grand, peu importe par exemple que la formule comporte n ou (n - 1). Comme l'écrivait en bougonnant K. Pearson à Gosset : "il n'y a que les méchants brasseurs pour prendre n si petit que la différence n'est pas de l'ordre des erreurs !" (cité d'après E.S. Pearson *Biometrika* T. 30, & *Studies* p. 368). Cependant certains auteurs ont entrepris d'affranchir la statistique et notamment celle des petits échantillons de l'hypothèse de normalité et plus généralement de toute hypothèse relative à la forme des lois : c'est la statistique non paramétrique que nous expliquons sur un exemple simple : trente élèves, 15 de la classe A, 15 de la classe B composent ensemble : les dix premiers appartiennent à la classe A; ce résultat en faveur de A est-il (sous réserve d'une correction équitable) compatible avec l'hypothèse que A et B sont de même force ? On voit que si les trente élèves sont rangés au hasard, la probabilité est seulement :

$$2 \times (15!/(10! \times 5!))/(30!/(20! \times 10!)) \approx 2 \times 10^{-4}$$

que l'une des classes emporte les 10 premières places : on conclut donc que le classement observé n'est pas dû au hasard. Dans la 6° édition de son *Design of Experiments*, Fisher tout en revendiquant la priorité dans la conception de ces épreuves ordinales, en condamne l'emploi généralisé; il affirme sa confiance dans les épreuves fondées sur la loi normale et souligne qu'il n'est pas sans danger de se priver de conclure, à force de raisonner comme si on était plus ignorant qu'on ne l'est en réalité (comme si on ignorait tout des valeurs numériques, de la forme des distributions; et ne connaissait que des rangs). Quant à nous, nous soulignerons que l'usage des méthodes ordinales apparaît difficile dans l'analyse des nuages de points multidimensionnels (cf § 3); mais la statistique non paramétrique a pu inspirer certaines méthodes de traitement des données non numériques (cf § 2.5) notamment l'analyse des proximités de R.N. Shepard (cf § 3.4.1); et elle peut aider à éprouver la validité des particularités observées sur un graphique issu d'une analyse factorielle.

2.3.3. Estimation par le maximum de vraisemblance : Nous avons rencontré le problème de l'estimation de la variance d'une loi normale d'après un échantillon $\{x^i\}$. Plus généralement on pourra considérer l'hypothèse qu'une grandeur aléatoire $x \in X$ (x peut être un nombre, $x \in R$; ou un vecteur, e.g. un système de mensurations effectuées sur un végétal d'une espèce donnée : $x = \{x_1, \ldots, x_s\} \in R^s = X$) suit une loi de la forme : $p(\theta^1, \ldots, \theta^r; x_1, \ldots, x_s)\, dx_1 \ldots dx_s$; où $\theta^1, \ldots, \theta^r$ sont des paramètres qu'il reste à préciser (e.g. si p est une loi normale de dimension s, les paramètres θ^j seront les s composantes de la moyenne, et les $s \times (s + 1)/2$ variances et covariances); en bref nous écrirons cette loi p_X^θ, ou $p_X^\theta(x)dx$; et noterons Θ l'espace où varie le paramètre multidimensionnel $\theta (\theta = \{\theta^1, \ldots, \theta^r\})$. Etant donné un échantillon $y = x^I = \{x^1, \ldots, x^n\} \in X^n = Y$, il faut trouver $\theta \in \Theta$: tel est le problème fondamental de l'estimation en statistique paramétrique. Même si, comme nous, on conteste que la loi de x soit vraiment de la forme p_X^θ, on peut parfois accepter d'associer à un échantillon empirique y, une loi p_X^θ qui en soit somme le résumé ...

Voici comment K. Pearson (dont nous avons dit l'amour pour les grands échantillons; amour satisfait par le labeur de Bénédictin, de son laboratoire) considère le cas unidimensionnel : X = R. A l'échantillon y, d'effectif élevé, est associé son histogramme, qui représente approximativement la courbe de la fonction de densité cherchée, $p_X^\theta(x)$. L'estimation des paramètres ($\theta = \{\theta^1, \ldots, \theta^r\}$) se fait par ajustement de la fonction à l'histogramme; un crière d'ajustement peut être fourni par la méthode des moindres carrés; mais K. Pearson aboutit bien plus rapidement à un θ satisfaisant par sa méthode des moments (cf On the Systematic Fitting of Curves; *Biometrika* T. 1, 1901-2). Le *moment* d'ordre n d'une distribution de densité $p_X^\theta(x)$ est l'intégrale $\int p_X^\theta(x)\, x^n dx$; Pearson choisit $\theta = \{\theta^1, \ldots, \theta^r\}$ en sorte que les r premiers moments de la loi p_X^θ et de l'histogramme coïncident (au passage il traite de la correction requise sur les moments de l'histogramme, du fait que les x^i de l'échantillon ont été groupés en classes auxquelles le dessin de l'histogramme attribue indûment une densité uniforme : c'est la correction d'abord proposée par Shepard).

En pratique, la méthode des moments est raisonnable. En théorie, rien ne permet de démontrer qu'elle soit optimale en aucun sens du terme. Dès le premier article de son oeuvre si féconde, (On an absolute criterion for fitting frequency curves; *Messenger of Mathematics*; T. 41, 1912; et *Collected papers*, T. 1), R.A. Fisher, encore étudiant à Cambridge, propose le critère appelé dans la suite critère du maximum de vraisemblance : l'échantillon $y = \{x^1, \ldots, x^i, \ldots, x^n\}$ étant donné, choisir θ (paramètre, multidimensionnel en général) rendant maximum le produit des densités :

$$\Pi\{p_X^\theta(x^i) \mid i = 1, \ldots, n\} = p_X^\theta(x^1) \times \ldots \times p_X^\theta(x^i) \times \ldots \times p_X^\theta(x^n).$$

Ce problème mit entre Pearson et Fisher une opposition que le temps ne put réduire : l'année même de sa mort, K. Pearson publie dans *Biometrika* (T. 28, 1936) "Method of Moments and Method of Maximum Likelihood"; à quoi R.A. Fisher répond après la mort de Pearson, par "Professor Karl Pearson and the Method of Moments" (in *Annals of Eugenics* T. 7, 1937; reproduit dans *Contributions to Mathematical Statistics*, Wiley, N.-Y. 1950). Nous ne tenterons pas d'analyser puis d'arbitrer cette controverse : il suffira ici de rappeler quelques principes.

Fisher refuse de dire que la valeur de θ rendant maximum le produit $\Pi\{p_X^\theta x^i)\}$ est la valeur la plus probable parmi celles conduisant à l'échantillon $y = \{x^i\}$: car pour assigner aux θ une probabilité *a posteriori* (d'après y) il faudrait partir d'une distribution de probabilité *a priori* $\pi(\theta)d\theta$, et raisonner suivant le principe de probabilité des causes (cf § 1.4.2); et Fisher ne postule pas de probabilité *a priori* (conception dite *non Bayesienne*). Mais puisque θ est le paramètre qui conduit avec la plus forte probabilité à l'échantillon y (plus précisément, est maxima pour θ la densité de probabilité en y), Fisher, propose d'appeler θ, paramètre le plus vraisemblable. (Sur l'origine de ce principe chez D. Bernoulli et Gauss cf § 1.5.3). Schématiquement, on dira que l'hypothèse paramétrique que x suit une loi p_X^θ avec $\theta \in \Theta$, définit une application généralisée (dite transition probabiliste, cf § 3.5.1) p_X^Θ de Θ vers X; application qui à $\theta \in \Theta$ fait correspondre non une image ponctuelle unique x, mais une valeur étalée dans X suivant la loi p_X^θ; de même on a l'application p_Y^Θ de Θ vers $Y = X^n$, qui à $\theta \in \Theta$ fait correspondre la loi $p_Y^\theta = (p_X^\theta)^n$, de l'échantillon y d'effectif n. Résoudre le problème de l'estimation, c'est en quelque sorte inverser l'application p_Y^Θ par un estimateur ε_Θ^Y application (déterministe, univalente, ordinaire) de Y dans Θ, qui à l'échantillon y, associe le paramètre estimé $\varepsilon(y) = \theta$. Il est clair que l'application composée $\tau_\Theta^\Theta = \varepsilon_\Theta^Y \circ p_Y^\Theta$ est elle-même une application généralisée (transition probabiliste) : p fait passer de θ à la distribution p_Y^Θ sur Y, et ε ramène cette distribution de Y vers Θ, suivant une autre distribution que l'on souhaite être concentrée autour du θ de départ, mais qui ne peut être réduite à ce point (cf fig. 2-5).

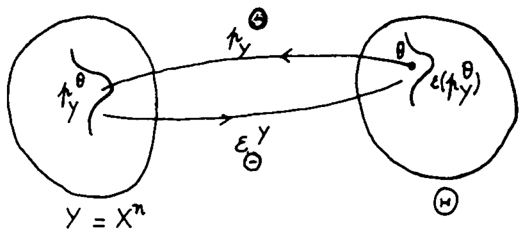

Figure 2-5 : Schéma de l'estimation ; on a symbolisé par des courbes en cloche la distribution p_Y^θ, et son image de retour $\varepsilon(p_Y^\theta)$ par l'estimation ε.

Un autre schéma géométrique suggère l'estimateur fourni par le principe du maximum de vraisemblance et ses propriétés asymptotiques quand n (effectif de l'échantillon) tend vers l'infini. Soit $\mathcal{P}_X$ l'espace des lois de probabilité sur X; dans $\mathcal{P}_X$, à l'hypothèse paramétrique p_X^Θ correspond une sous variété $\mathcal{L}$ de dimension r (nombre des composantes de $\theta = (\theta^1, \ldots, \theta^r) \in \Theta \subset R^r)$: $\mathcal{L} = \{p_X^\theta | \theta \in \Theta\} \subset \mathcal{P}_X$. A un échantillon $y \in X^n$, correspond sa loi de fréquence $f_X(y) \in \mathcal{P}_X$ (la notion de loi de fréquence est très claire si X est un ensemble fini I : la fréquence f_I de i est le nombre n_i de fois que i figure dans l'échantillon y, divisé par n : $f_i = n_i/n$; dans le cas continu, nous ne tenterons pas de préciser; cf [Dis χ^2 Loi] TII B n° 4); faire une estimation de θ, c'est associer à f_X une projection p_X^θ sur $\mathcal{L}$. Il se trouve que la méthode du maximum de vraisemblance est une sorte de projection orthogonale (pour la métrique du χ^2, cf § 3.2) sur $\mathcal{L}$ dans $\mathcal{P}_X$ (cf fig. 2-6): la distribution $f_X(y)$, loi de fréquence de l'échantillon y, se projette sur $\mathcal{L}$, en $p_X^{\varepsilon(y)}$, loi dont le paramètre $\theta = \varepsilon(y)$ est celui estimé à partir de y par le principe du maximum de vraisemblance. A une loi p_X^θ il correspond des échantillons, y_θ dont les lois de fréquence $f_X(y)$ sont distribuées dans $\mathcal{P}_X$ autour de p_X^θ; et par projection sur $\mathcal{L}$, on a autour de p_X^θ une distribution des valeurs de θ estimées par ε à partir des échantillons issus de p_X^θ; c'est ce qu'on a noté $\varepsilon(p_Y^\theta)$ sur la figure 2-5. Pour n grand, il résulte du théorème centrel limite (cf § 1.5.3) que cette distribution tend à être normale.

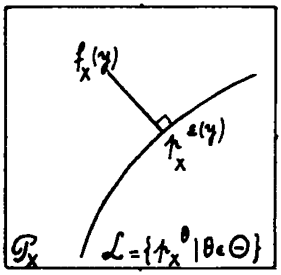

Figure 2-6 : L'estimation du maximum de vraisemblance est une sorte de projection orthogonale.

Ce n'est pas le lieu, dans une introduction historique à l'analyse des données, d'exposer la théorie de l'estimation de Fisher (dont nous suggérerons seulement le programme par quelques titres : dispersion de $\varepsilon(p_Y^\theta)$ (*); calcul de l'information de Fisher, cf [Dis χ^2 loi], TII B n° 4; cas des petits échantillons et cas limite $n \to \infty$; théorie de la probabilité fiduciaire, ou essai d'associer à un échantillon y, sans postuler de probabilité *a priori* dans un cadre Bayesien, une loi de probabilité *a posteriori* φ_Θ^Y du paramètre θ; mais ce dernier essai n'a pas été généralement accepté) et les recherches qu'elle a inspirées (notamment à J. Neyman et E.G. Pearson - le fils de K. P. - sur les épreuves de validité ou *tests* des hypothèses; cf *infra* § 2.3.6). Il nous importait d'en évoquer, dans le langage qui nous est familier, l'ampleur géométrique.

Quant aux mérites pratiques, nous serons réservés : l'optimisation des estimateurs ne vaut que sous une hypothèse paramétrique rarement solide; mêlez à une loi normale 1 % d'individus distribués suivant la loi de Cauchy $dx/(\pi(1 + x^2))$ (loi si bien étudiée par Fisher; cf e.g. Theory of statistical estimation; in *Proc. of The Cambridge Phil. Soc.* T. 22; 1925 et *Collected Papers*, ou *Contributions*) et l'estimateur de la tendance centrale par la moyenne de l'échantillon perd sa valeur : on songe à utiliser la médiane, à moins d'avoir éliminé de quelque manière les étrangers les plus éloignés du centre (sur ce problème de l'*estimation robuste*, i.e. à l'épreuve des perturbations affectant les hypothèses paramétriques, citons une revue de P.J. Hubert : Robust statistics : a review, in *The Ann. of Math. Stat.* T. 42, 1972; où l'on trouve notamment cette règle d'estimation de la moyenne proposée par Gastwirth : $0{,}3\ x_{1/3} + 0{,}4\ x_{1/2} + 0{,}3\ x_{2/3}$ - où x_p désigne la valeur qui est précédée par une fraction p de l'échantillon, e.g. $x_{1/2}$ est la médiane). Mathématiquement, la conception de la statistique paramétrique non-bayésienne est parfaitement claire; philosophiquement, elle résulte d'un compromis quant au rôle de l'*a priori*. D'une part, rigidité (a priorisme) pour la forme postulée de la loi p_X^θ; d'autre part liberté absolue pour la distribution de θ (sans loi *a priori* π_θ). Depuis vingt ans on s'est dans l'un et l'autre sens écarté de la norme Fisherienne; nous avons cité plus haut la statistique non-paramétrique (préfigurée chez Fisher mais dont l'emploi généralisé est refusé par lui;

(*) *En appendice au deuxième article de cette série historique (II La Biométrie; Cahier n° 2) une note due à L. Lebart traite d'une inégalité bornant inférieurement la variance d'un estimateur; cette inégalité, connue sous les noms de Fréchet-Darmois-Cramer-Rao, est au terme d'une suite de travaux à l'origine desquels on trouve R. Fisher.*

cf § 2.3.2); il faut ajouter que les méthodes néo-bayesiennes sont en faveur, notamment dans le domaine multidimensionnel en discrimination, en régression; où, la forme des distributions étant totalement inconnue *a priori*, l'échantillon des données (nuage de points dans R^n) doit être accepté comme une sorte de distribution empirique (n'est-ce pas là le principe général de l'analyse des données telle que nous la pratiquons, cf § 3; et en particulier de la régression et discrimination Bayesienne, après analyse factorielle; cf § 3.8.2).

2.3.4. Analyse de la variance et plan d'expérience : Comme le nom l'indique, l'analyse de la variance consiste en une décomposition de la variance, ou inertie d'un nuage de points. De telles décompositions, le théorème de Huyghens et les changements orthogonaux de coordonnées offrent des exemples. Mais leur usage en statistique avec le dénombrement des degrés de liberté afférentes à diverses hypothèses et l'isolement des résidus aléatoires, appartient croyons-nous à Fisher; et Mather et Yates dans l'introduction à l'édition des *Collected Papers* font remonter à la communication déjà citée du Congrés de Toronto en 1924, l'usage systématique de ce qu'on appelle aujourd'hui analyse de variance.

De même l'expérimentation systématique en butte à des erreurs aléatoires dont l'amplitude peut non seulement perturber mais masquer les effets à observer, avait avant Fisher, préoccupé les agronomes. Mais Fisher a vu le premier qu'une analyse de la variance, en vue de laquelle doit être conçue l'expérimentation, permet en mesurant l'importance relative de diverses causes, d'éprouver des hypothèses relatives à celles-ci. C'est la doctrine du *plan d'expérience*, oeuvre d'un géomètre devenu en 1919 le statisticien du Domaine d'expériences agronomiques de Rothamsted.

Nous nous bornerons ici à expliquer l'analyse de variance sur deux cas; dont l'un nous donnera aussi un modèle assez simple mais non trivial, de plan d'expérience. Au § 2.3.6 nous reprendrons la doctrine de Fisher sur l'expérimentation et l'induction.

De façon précise considérons un ensemble x^I, indicé par l'ensemble fini I, de mesures x^i d'une même grandeur : $x^I = \{x^i | i \in I\} \in R^I$: par exemple I est un ensemble de parcelles de terre dont chacune a reçu un traitement propre; et x^i est la production de la parcelle i (le cas de grandeurs multidimensionnelles est analogue : il requiert seulement un indice supplémentaire j pour distinguer les diverses mesures x^{ij} afférentes à un cas i; cf § 2.3.5). Faisons choix d'une décomposition de R^I en somme directe de sous-espaces deux à deux orthogonaux (dans R^I muni de la métrique usuelle : $\|x^I\|^2 = \Sigma\{(x^i)^2 | i \in I\}$). Notons $\{X^a | a \in A\}$ la famille indicée par $a \in A$, de ces sous-espaces; tout vecteur x^I de R^I s'écrit de manière unique comme une somme de composantes x^a chacune appartenant à l'un des axes X^a; on a :

$$R^I = \oplus \{X^a | a \in A\};$$

$$x^I = \Sigma\{x^a | a \in A\};$$

$$\|x^I\|^2 = \Sigma\{\|x^a\|^2 | a \in A\}.$$

Au § 2.3.2 on a considéré la décomposition de R^I en deux sous-espaces Δ et H; ce qu'on peut récrire dans les notations plus générales posées ici :

$$A = \{\Delta , H\}; \quad X^\Delta = \Delta; \quad X^H = H,$$

$$X^\Delta = \{x^I | x^I \in R^I;\ \forall\ i,i' \in I : x^i = x^{i'}\};$$
$$X^H = \{x^I | x^I \in R^I;\ \Sigma\{x^i | i \in I\} = 0\};$$

i.e. la diagonale X^Δ est le lieu des points x^I correspondant à un ensemble de mesures toutes égales; l'hyperplan X^H supplémentaire de la diagonale, est le lieu des points x^I correspondant à un ensemble de mesures dont la moyenne est nulle;

$x^\Delta = \{\delta^i \bar{x} | i \in I\}$, point dont toutes les coordonnées sont égales à la moyenne $\bar{x}$;

$x^H = \{(x^i - \bar{x}) | i \in I\}$, point dont toutes les coordonnées sont rapportées à la moyenne $\bar{x}$;

$\|x^\Delta\|^2 = n(\bar{x}^2) = (n^{1/2}\ \bar{x})^2$;

$\|x^H\|^2 = \Sigma\{(x^i - \bar{x})^2 | i \in I\}$; c'est-à-dire, à un coefficient (1/n) près, la variance de l'ensemble x^I de mesures indicé par i (exactement, le moment centré d'ordre 2 de cet ensemble, si chaque mesure reçoit poids 1).

Les décompositions orthogonales de R^I considérées en statistique comprennent toujours comme premier sous-espace l'axe Δ; les autres sous-espaces fournissent donc une décomposition orthogonale de H. Nous noterons :

$A = \{\Delta\} \cup A_H$; $A_H = A - \{\Delta\}$;

$X^H = H = \oplus\ \{X^a | a \in A_H\}$;

$x^H = \Sigma\{x^a | a \in A_H\}$;

$\|x^H\| = \Sigma\{\|x^a\|^2 | a \in A_H\}$.

Cette dernière formule fournit une décomposition de $\|x^H\|^2$; donc une décomposition, une analyse de la variance. Avant d'expliquer l'usage fait de cette analyse dans le cadre de l'hypothèse de normalité, donnons deux exemples simples de telle décomposition correspondant à une structure convenable de I.

Exemple I : variance intraclasse et variance interclasse. L'ensemble I est muni d'une partition en un ensemble Q de classes non-vides et deux à deux disjointes :

$I = \cup\ \{q | q \in Q\}$;

$\forall\ q,\ q' \in Q : (q \cap q' = \emptyset) \Leftrightarrow (q \neq q')$;

on note q(i) la classe de Q à laquelle appartient l'élément i.
On décompose R^I en une somme de trois sous-espaces X^Δ, X^Q, X^{I-Q}, définis comme suit :

$X^\Delta = \Delta$, la diagonale Δ usuelle;

$X^Q = \{x^I | x^I \in H;\ \forall\ q \in Q,\ \forall\ i,\ i' \in q : x^i = x^{i'}\}$; i.e. X^Q est le lieu des points x^I appartenant à H (systèmes de mesures de moyenne nulle) et tels que la mesure x^i ne dépend que de la classe q de i; on a dim X^Q = card Q - 1.

$x^{I-Q} = \{x^I | x^I \in R^I;\ \forall q \in Q : \Sigma\{x^i | i \in q\} = 0\}$; i.e. X^{I-Q} est le lieu des points x^I représentant un système de mesures x^i ayant moyenne nulle sur chaque q; dim X^{I-Q} = card I - Card Q.

Pour définir les composantes x^Q et x^{I-Q}, il est commode de noter : $\bar{x}^q$ la moyenne des x^i d'une classe $q(\bar{x}^q = (1/\text{Card } q)\ \Sigma\{x^i | i \in q\})$; on a alors :

$$x^Q = \{(\bar{x}^{q(i)} - \bar{x}) | i \in I\} \in X^Q \subset R^I$$
$$x^{I-Q} = \{(x^i - \bar{x}^{q(i)}) | i \in I\} \in x^{I-Q} \subset R^I$$

(i.e. dans x^{I-Q}, chaque mesure est diminuée de la moyenne de sa classe; dans x^Q, chaque mesure est remplacée par la moyenne de sa classe, diminuée de la moyenne générale). Ceci posé, on a la formule de décomposition :

$$\|x^H\|^2 = \|x^Q\|^2 + \|x^{I-Q}\|^2;$$
$$\|x^Q\|^2 = \Sigma\{(\bar{x}^{q(i)} - \bar{x})^2 | i \in I\}$$
$$= \Sigma\{\text{Card } q\ (\bar{x}^q - \bar{x})^2 | q \in Q\};$$
$$\|x^{I-Q}\|^2 = \Sigma\{(x^i - \bar{x}^{q(i)})^2 | i \in I\}$$
$$= \Sigma\{\Sigma\{(x^i - \bar{x}^{q(i)})^2 | i \in q\} | q \in Q\};$$

la somme partielle $(1/n)\|x^Q\|^2$ est appelée variance interclasse, ou variance entre les classes (exactement : variance entre les moyennes des classes, chacune recevant pour poids son effectif Card q); la somme partielle $(1/n)\|x^{I-Q}\|^2$ est appelée variance intraclasse ou variance intérieure aux classes (chaque mesure est rapportée à la moyenne de sa classe; on obtient la moyenne des variances des classes, chacune pondérée par Card q). Au § 2.3.5 nous considérerons la généralisation de cet exemple au cas multidimensionnel.

Exemple 2 : carré latin. L'ensemble I (des cas observés) a pour cardinal un carré r^2. Soient T_α, T_β, T_γ trois ensembles finis de cardinal r, qu'on pourra regarder comme des ensembles de traitements ou de conditions dans chacun desquels on doit choisir une modalité pour chaque cas i : e.g. T_α ensemble de terrains, T_β ensemble d'engrais, T_γ ensemble de semences. On a trois applications t_α, t_β, t_γ de I dans les trois ensembles T_α, T_β, T_γ : $t_\beta(i)$ étant, e.g., l'engrais choisi pour la parcelle i. La structure de *carré latin* est définie par la condition que les trois applications produit $t_\alpha \times t_\beta$, $t_\beta \times t_\gamma$ et $t_\gamma \times t_\alpha$ mettent respectivement I en correspondance biunivoque avec $T_\alpha \times T_\beta$, $T_\beta \times T_\gamma$ et $T_\gamma \times T_\alpha$. Souvent, on donne de cette structure une présentation dissymétrique : I est un carré r x r dont l'ensemble des lignes est identifié à T_α et l'ensemble des colonnes à T_β : $I \approx T_\alpha \times T_\beta$; l'application t_γ est alors figurée en écrivant dans chaque case du carré I un élément de l'ensemble T_γ : la condition de carré latin est vérifiée si tout élément de T_γ est inscrit une fois et une seule sur chaque ligne (biunivocité de $t_\gamma \times t_\alpha$) et sur chaque colonne (biunivocité de $t_\beta \times t_\gamma$). Et dans l'expérimentation I est bel et bien un champ carré, divisé en r^2 parcelles : T_α et T_β sont les conditions de position (il faut tenir compte de l'hétérogénéité inévitable d'une terre; particulièrement si elle a plusieurs fois servi à des expériences de culture); et T_γ est un ensemble d'engrais, ou de semences etc ... $t_\gamma(i)$ étant choisi pour la parcelle i. La recherche des carrés latins est un très ancien problème d'analyse combinatoire; et l'usage de certains d'entre eux comme grille d'expérience se rencontre au XIX° siècle. Reste à faire l'analyse de variance.

Pour cela, on décompose R^I en une somme directe de cinq sous-espaces que nous décrirons brièvement : $A = \{\Delta, \alpha, \beta, \gamma, \rho\}$:

X^Δ : la diagonale usuelle :

$X^\alpha = \{x^I | x^I \in H; \forall\, i, i' \in I : (t_\alpha(i) = t_\alpha(i')) \Rightarrow x^i = x^{i'}\}$; i.e. la moyenne des x^i est nulle, et la coordonnées x^i ne dépend que du traitement (ou de la condition) $t_\alpha(i)$; $\dim X^\alpha = r - 1$

X^β, X^γ : comme X^α, *mutatis mutandis*.

X^ρ : supplémentaire orthogonal de $(X^\Delta \oplus X^\alpha \oplus X^\beta \oplus X^\gamma)$ dans R^I : c'est le résidu; la merveille dans cette décomposition est que les trois sous-espaces X^α, X^β, X^γ soient orthogonaux entre eux : ici intervient la structure de carré latin.

Venons maintenant aux applications statistiques de l'analyse de variance, applications qui généralisent l'étude faite ci-dessus pour la loi de Student (§ 2.3.2). Supposons que les x^i soient des variables aléatoires deux à deux indépendantes, distribuées normalement avec la même variance σ^2 et des moyennes m^i dont l'expérience a pour but de manifester les inégalités : par exemple le rendement x^i d'une parcelle i - qui d'ailleurs ne peut en toute rigueur être distribué normalement, car il est borné et non négatif - a une moyenne positive m^i qui dépend des traitements et conditions propres à la parcelle i, et de plus fluctue autour de cette moyenne suivant des aléats de même amplitude que ceux auxquels est en butte toute autre parcelle i'. Dans R^I (toujours muni de la métrique usuelle) x^I a une distribution normale sphérique de centre $m^I = \{m^i | i \in I\}$; chaque composante x^a de x^I a de même dans le sous-espace X^a, une distribution normale sphérique de centre m^a (composante de m^I dans X^a). Reportons-nous aux exemples 1 et 2 : m^Δ n'est autre que le point dont toutes les coordonnées sont égales à $\bar{m}$ (moyenne générale des m^i); m^Q (exemple 1) a pour i-ème coordonnée la moyenne $\bar{m}^{q(i)}$ de la classe q(i) diminuée de la moyenne générale $\bar{m}$; de même m^α (exemple 2; ou m^β, ou m^γ) a pour i-ème coordonnée $(\bar{m}^{t_\alpha(i)} - \bar{m})$ où $\bar{m}^{t_\alpha(i)}$ est la moyenne afférente aux individus soumis au même traitement $t_\alpha(i)$ que i; m^{I-Q} (exemple 1) a pour i-ème coordonnée la moyenne m^i diminuée de la moyenne $\bar{m}^{q(i)}$ propre à la classe q dans laquelle rentre .

i : c'est donc un reste : il en est de même de m^{ρ} dans l'exemple 2. Généralement pour l'expérimentateur m^{Δ} est certainement non nul; tandis que m^{I-Q} ou éventuellement m^{ρ} (le reste, le résidu) est certainement nul; la question est de savoir si les autres composantes de m^{I} sont ou non, nulles : dans l'exemple 1, $m^{Q} = 0$ signifie qu'il n'y a pas de différence systématique entre les classes (qui e.g. correspondent chacune à une espèce); dans l'exemple 2, $m^{\gamma} = 0$ signifie que le traitement T_{γ} est sans influence etc. Par l'analyse de variance, le statisticien contribue à résoudre cette question. En effet chacune des composantes $\|x^{a} - m^{a}\|^2$ est distribuée comme un χ^2 dont la dimension n^{a} est celle du sous-espace X^{a} : donc dans l'exemple 1, si $m^{Q} = 0$ (on a déjà dit que $m^{I-Q} = 0$) le rapport $\|x^{Q}\|^2/\|x^{I-Q}\|^2$ de la variance entre classes à la variance intérieure aux classes (dite encore variance résiduelle) est distribuée comme un rapport de deux χ^2 ayant respectivement pour dimension (Card Q-1) et (Card I - Card Q); dans l'exemple 2, si $m^{\gamma} = 0$ (on a déjà dit que $m^{\rho} = 0$) le rapport $\|x^{\gamma}\|^2/\|x^{\rho}\|^2$ de la variance entre modalités du traitement T^{γ} à la variance résiduelle est le rapport de deux χ^2 de dimensions respectives $(r - 1)$ et $(r^2 - 3r + 2)$; (car $\dim R^{I} = \text{Card } I = r^2$; $\dim X^{\alpha} = \dim X^{\beta} = \dim X^{\gamma} = r - 1$; $\dim \Delta = 1$; le reste est pour X^{ρ}, soit : $r^2 - (3(r - 1)) - 1 = r^2 - 3r + 2$). Or la loi d'un rapport de deux χ^2 a été étudiée par Fisher : si le rapport trouvé pour $\|x^{Q}\|^2/\|x^{I-Q}\|^2$ a une faible chance d'être atteint ou dépassé, on conclut au rejet de l'hypothèse $m^{Q} = 0$. De même dans l'exemple du carré latin si $\|x^{\gamma}\|^2/\|x^{\rho}\|^2$ est trop élevé, c'est que $m^{\gamma} \neq 0$ (le traitement agit sensiblement) : on comprend ici pourquoi deux ensembles T^{α} et T^{β} peuvent être consacrés à l'influence des bandes (lignes ou colonnes) d'un champ sur le rendement, alors qu'évidemment l'hétérogénéité du champ n'est pas l'objet de l'expérimentation : c'est que la combinatoire judicieuse du plan d'expérience permet d'isoler une variance résiduelle $\|x^{\rho}\|^2$ beaucoup plus faible que si l'on n'avait pas explicitement pris en compte les coordonnées de la parcelle i; et ainsi on mettra plus facilement en évidence comme significatives des variances $\|x^{\gamma}\|^2$ dues au traitement, par rapport à $\|x^{\rho}\|^2$ qui nous sert d'étalon de variance.

L'analyse de variance de l'exemple 1 vaut pour tout ensemble I de mesures réparties en classes d'effectifs quelconques. La structure de carré latin au contraire n'appartient évidemment pas à un ensemble d'observations recueillies sans préparation : elle requiert un plan d'expérience. On en peut concevoir de plus simples (étant donné n blocs, chacun subdivisé en r parcelles, et un ensemble de r traitements; appliquer à chaque parcelle un traitement choisi au hasard, sous la condition que chaque traitement soit appliqué à une parcelle et une seule de chaque bloc) et de plus complexes (pour étudier, autant qu'il est possible, sur un ensemble limité de cas les diverses combinaisons des modalités de plusieurs ensembles de traitements). Nous avons choisi l'exemple du carré latin parce que sans être très complexe, il montre qu'une structure combinatoire peut donner des propriétés d'orthogonalité (dans l'ex. 2, entre X^{α}, X^{β}, X^{γ}) requises par l'analyse de variance, et non vérifiées sur une grille de combinaison quelconque des conditions et des traitements.

On peut maintenant suggérer ce qu'est le plan d'expérience à la Fisher, conçu pour une analyse de variance des données. Le plan prend explicitement en compte un certain nombre de conditions ou de traitements α, β, γ ... (e.g. semence, engrais, arrosage ...), ayant respectivement pour ensemble de modalités T_{α}, T_{β}, T_{γ}, ...

L'ensemble I des cas d'expérience est muni d'applications t_{α}, t_{β}, t_{γ}, ... dans T_{α}, T_{β}, T_{γ}, ..., qui attribuent à chaque cas i des modalités $t_{\alpha}(i)$, $t_{\beta}(i)$, $t_{\gamma}(i)$,... (choix de la semence; de l'engrais; du mode d'arrosage). Les applications t_{α}, t_{β}, t_{γ}, ... doivent pour assurer l'orthogonalité des sous-espaces de R^I associés à divers modèles d'action et d'interaction des traitements, satisfaire à des conditions qui posent souvent de très difficiles problèmes d'analyse combinatoire. Si toutefois une solution est trouvée, on en déduit sans peine un grand nombre d'autres par permutation aléatoire, notamment en composant les t_{α}, t_{β}, t_{γ}, ..., avec des permutations σ_{α}, σ_{β}, σ_{γ} des ensembles T_{α}, T_{β}, T_{γ}, ...; et posant $t'_{\alpha} = \sigma_{\alpha} \circ t_{\alpha}$, $t'_{\beta} = \sigma_{\beta} \circ t_{\beta}$, $t'_{\gamma} = \sigma_{\gamma} \circ t_{\gamma}$, ... : par exemple, un seul carré latin r x r peut ainsi être soumis à $(r!)^3$ transformations différentes (qui il est vrai, ne donnent pas des combinaisons toutes différentes entre elles). Il est essentiel pour éviter tout effet systématique non voulu, de n'appliquer le plan d'expérience qu'on a choisi, qu'avec permutation aléatoire : c'est ce que Fisher appelle *randomisation*.

Soulignons que le but d'un tel système d'expériences élémentaires n'est pas seulement de recueillir des informations sur une grande variété de cas couvrant judicieusement l'ensemble du domaine que l'on entend explorer, c'est aussi de permettre à l'analyse de variance de décider quels effets sont significatifs en rapportant ceux-ci à l'étalon fourni par la variance résiduelle. R.A. Fisher l'explique ainsi (*the Design of Experiments* § 28) : "Le principe essentiel qui régit notre analyse est que les erreurs dues à l'hétérogénéité du sol seront dans une expérience bien conçue, séparées en deux parts. La première, aussi grande que possible, sera complètement éliminée, grâce à l'arrangement adopté, des comparaisons expérimentales; et disparaîtra dans les calculs du laboratoire de statistique, de l'estimation de l'erreur. Quant au reste qui ne peut être ainsi traité, on ne tentera pas de l'éliminer sur le terrain, mais au contraire méticuleusement soumis à une permutation aléatoire [*randomised*, écrit Fisher], il fournira une estimation valable des erreurs auxquelles l'expérience est en butte".

D'après l'exemple du carré latin, on concevra que Fisher puisse en général, d'une part éliminer (complètement, affirme-t-il : ce qui n'est strictement vrai que des effets additifs prévus dans le modèle) les effets de certaines conditions explicitement comprises dans la combinatoire du plan (e.g. effets de bande des lignes et colonnes); et d'autre part calculer la variance résiduelle étalon requise pour les comparaisons.

Inutile de répéter qu'il est impossible de condenser en un § un traité du plan d'expérience. Nous voulons seulement d'une part faire admirer au spécialiste de l'analyse des données la puissance des conceptions géométriques de Fisher (conceptions qui valent pour nous, car l'analyse factorielle (*) et la classification automatique sont fondées sur des décompositions de la variance - inertie - d'un nuage de points; cf §§ 3.8.3-3.8.4); et d'autre part préparer une confrontation entre la méthode fishérienne d'induction fondée sur l'expérimentation planifiée et l'analyse de variance (**), et une autre méthode - la seule possible croyons-nous dans bien des disciplines - fondée sur l'analyse (il faudrait dire la synthèse) d'observations multidimensionnelles (§ 2.3.6).

2.3.5. *Analyse discriminante* : sur une décomposition de la variance se fonde aussi l'analyse discriminante telle que la conçoit Fisher. C'est une généralisation au cas multidimensionnel de l'exemple 1 du § 2.3.4 : rappelons ce problème (déjà considéré par K. Pearson) avec les notations que nous avons posées ailleurs ([Sép. Corr.]). Soit I un ensemble d'individus répartis en un ensemble Q de classes et chacun décrit par un système J de mesures. On note $i^J = \{i^j | j \in J\} \in R^J$, le vecteur de description de l'individu i : ainsi i^j désigne la j-ème mesure effectuée sur l'individu i. Le problème fondamental de l'analyse discriminante est d'affecter dans R^J à chaque classe q de Q, un domaine de forme simple contenant les individus de cette classe; de telle sorte que les mesures j fournissent une définition des classes; et que tout individu nouveau s puisse d'après son vecteur de description s^J, être affecté à la classe q dans le domaine de laquelle il tombe (par exemple, on déterminera l'espèce d'un animal d'après des mesures prises sur son crâne). Comme l'a clairement marqué Fisher, le problème de la discrimination rentre dans celui de la régression : les mesures i^j sont les variables explicatives; la classe q(i) de l'individu i est la variable à expliquer, variable qui prend ses valeurs non dans la droite R, mais dans l'ensemble fini Q : on cherche, en quelque manière, à exprimer q(i) en fonction des i^j (non sans une certaine imprécision; c'est-à-dire, ici, un certain taux d'erreurs).

Identifions dans nos formules chaque classe q avec l'ensemble des individus de I qui lui appartiennent; alors Q est un ensemble de parties de I et on a : $I = \cup \{q | q \in Q\}$, (les parties q de I étant non-vides et deux à deux disjointes). Bien que les individus i reçoivent ordinairement des poids égaux on note explicitement μ_i la masse attribuée à i; et on a pour I et les classes q les centres de gravité g^J et q^J :

$$M = \mu(I) = \Sigma\{\mu_i | i \in I\}; \quad \mu_q = \Sigma\{\mu_i | i \in q\};$$

(*) *Cyril Burt (in* L'analyse factorielle et ses applications; *Colloque C.N.R.S. 1955) : évoquant de fructueux échanges avec R.A. Fisher, rappelle que c'est à l'exemple de l'analyse de variance qu'il entreprit de faire l'analyse factorielle non sur une matrice de coefficients de corrélation, mais sur une matrice de covariance; et il ajoute - sans toutefois argumenter en détail - "Les analogies entre l'analyse factorielle et l'analyse de la variance sont évidentes".*

(**) *Méthode qui a connu une extension d'autant plus grande que les traitements et conditions α, β, γ, ..., ont dans le plan d'expérience des ensembles finis de modalités; ce qui permet de considérer des variables qualitatives dont le rôle est essentiel dans les sciences humaines notamment en psychologie. Mais cette concentration des mesures sur quelques modalités répétées, a l'inconvénient de réduire les capacités exploratrices du plan !; cf § 2.3.6.*

$$g^J = (1/M)\ \Sigma\{\mu_i\ i^J | i \in I\};\ q^J = (1/\mu_q)\ \Sigma\{\mu_i\ i^J | i \in q\}.$$

Ceci posé, pour une forme linéaire sur R^J : $u(x) = \Sigma\{u_j x^i | j \in J\}$, (ou : $v(x) = \Sigma\{v_j\ x^j | j \in J\}$), combinaison des mesures primaires, on définit la moyenne générale sur I (qui n'est autre que la valeur au point g^J), la moyenne sur la classe q (valeur au point q^J); et pour deux formes u(x) et v(x), plusieurs notions de covariance :

covariance globale, $\sigma(I)$:

$$\sigma(I)(u,v) = (1/M)\ \Sigma\{\mu_i\ u(i^J - g^J)\ v(i^J - g^J) | i \in I\};$$

covariance entre les classes, ou interclasse $\sigma(Q)$:

$$\sigma(Q)(u,v) = (1/M)\ \Sigma\{\mu_q u(q^J - g^J)\ v(q^J - g^J) | q \in Q\};$$

covariance intérieure aux classes, ou intraclasse, $\sigma(I - Q)$, (chaque individu étant rapporté au centre de sa classe q(i)) :

$$\sigma(I - Q)(u,v) = (1/M)\ \Sigma\{\Sigma\{\mu_i\ u(i^J - q^J)\cdot v(i^J - q^J) | i \in q\} | q \in Q\}.$$

La covariance globale est somme de la covariance interclasse et de la covariance intraclasse $\sigma(I) = \sigma(Q) + \sigma(I - Q)$. Une forme linéaire u telle que $\sigma(I)(u,u) = \sigma(Q)(u,u)$, et donc par différence $\sigma(I-Q)(u,u)=0$, a variance nulle sur chaque classe q et, par conséquent, y est constante : si elle ne prend pas même valeur aux centres de deux classes distinctes, u résoudra sans erreur le problème de la discrimination : la classe de i étant l'unique q telle que $u(q^J) = u(i^J)$. Une telle forme n'existe pas en général; mais à défaut, on cherche u rendant maximum le rapport rap = $\sigma(Q)(u,u)/\sigma(I)(u,u)$ de la variance entre les classes à la variance totale; (il est équivalent de demander que u rende maximum le rapport rap' = $\sigma(I)(u,u)/\sigma(I - Q)(u,u)$ de la variance totale à la variance intraclasse) car rap' = 1/(1 - rap); ou encore le rapport rap" = $\sigma(Q)(u,u)/\sigma(I - Q)(u,u)$, car rap" = rap' - 1).

S'il n'y a que deux classes, les hyperplans de niveau de la forme linéaire u réalisant ce maximum sont perpendiculaires à la droite joignant les centres de gravité des deux classes, dans la géométrie euclidienne pour laquelle la variance globale $\sigma(I)$ est sphérique (i.e; : I a même variance dans toute direction). Dans cette géométrie l'hyperplan médiateur du segment joignant les centres des classes constitue entre celles-ci une séparation intéressante, appelée hyperplan de Fisher : on affectera pour domaine à chaque classe dans R^J le demi-espace contenant son centre et ayant pour frontière l'hyperplan de Fisher. Il est équivalent de construire cet hyperplan médiateur dans la géométrie euclidienne pour laquelle est sphérique la variance intraclasse $\sigma(I-Q)$. Du point de vue de l'analyse des données, qui considère principalement les propriétés d'inertie des nuages de points, les diverses géométries se coordonnent harmonieusement : mais si l'on cherche dans les calculs de distance des épreuves de validité, il faut choisir une formule déterminée; d'où des controverses que nous ne détaillerons pas ici.

Car l'analyse discriminante pose au statisticien un problème de validité : une cloison (ou un système de cloisons s'il y a plus de deux classes) a été mise en place d'après des échantillons finis de points i des classes q (e.g. 20 crânes de l'espèce A; 30 de l'espèce B); entre ces échantillons mêmes, la séparation peut être imparfaite (des individus de A peuvent tomber dans le domaine de B; et *vice versa*); que vaut-elle pour les classes potentiellement infinies qu'on vise à séparer (pour un crâne quelconque de l'espèce A ou B) ? Si l'on accepte le modèle normal, on peut poser des problèmes mathématiques dont le premier est celui-ci, résolu par Hotelling : un nuage I de points de

R^J étant censé provenir d'une distribution normale centrée à l'origine (mais de matrice des variances-covariances inconnue), caractériser l'écart de cette distribution à l'origine : c'est à plusieurs dimensions (dans R^J) le problème résolu par Student pour une seule (sur la droite). Hotelling calcule une statistique T^2 (généralisant le t^2 de Student; cf § 2.3.2) qui n'est autre que le produit par (n-1) du carré de la distance du centre g^J à l'origine, dans la métrique euclidienne pour laquelle le nuage I a variance 1 dans toute direction; et il donne la distribution exacte de T^2, (sous l'hypothèse, répétons-le, que les i^J sont issus d'une loi normale centrée).

Nous ne tenterons pas de condenser les recherches de Fisher, ni celles - d'un même esprit - de l'école indienne (Mahalanobis, S.N. Roy); ni celles de A. Wald; signalons seulement que c'est à propos de la discrimination dans le cas des variables descriptives qualitatives, (et non quantitatives comme les i^J posées ci-dessus) que R. Fisher posa les équations de l'analyse des correspondances (cf § 3.5.2) sans toutefois leur demander autre chose qu'une fonction discriminante.

Depuis quelque vingt ans avec le développement du calcul électronique et les essais de reconnaissance des formes, l'analyse discriminante s'engage dans des voies nouvelles : sans hypothèse de normalité, le nuage des individus dans R^J est traité comme une distribution empirique (c'est le point de vue néo-bayesien; cf § 2.3.3 *in fine*). Certains auteurs, en discrimination comme en régression, procèdent en choisissant successivement pas à pas parmi l'ensemble J des variables descriptives, quelques unes de celles-ci qui apparaissent propres à fournir par combinaison linéaire une forme u(x) satisfaisante. Nous verrons au § 3.8.2 comment traiter par analyse de correspondance les problèmes de discrimination et de régression. Il importe toutefois de noter qu'il n'est pas toujours nécessaire de faire entrer dans l'analyse multidimensionnelle la partition Q à expliquer. Reprenons par exemple les données relatives au genre *Iris*, analysée par Fisher (The use of multiple measurements in taxonomic problems; in *Annals of Eugenics* T. 7 Pt. II, 1936; et *Contributions*). Pour chacune des trois espèces *I. Setosa*, *I. virginica* et *I. versicolor*, on a mesuré sur cinquante fleurs la largeur et la longueur des pétales et des sépales; soit un tableau 150 x 4 (150 fleurs x 4 mensurations; card J = 4). La fonction linéaire discriminante u déterminée par Fisher (comme rendant maximum $\sigma(Q)(u,u)/\sigma(I)(u,u)$; cf *supra*) sépare parfaitement les cinquante fleurs de *I. setosa* d'une part, et les cent fleurs de *I, versicolor* et *I. virginica* de l'autre; mais entre ces deux espèces il y a un certain empiètement. Y. Grelet et M. Roux ont repris ces données en soumettant d'abord le tableau 150 x 4 à l'analyse des correspondances (donc sans introduire l'information *a priori* que ces fleurs relèvent de trois espèces): le premier facteur obtenu a le même pouvoir discriminant que la fonction de Fisher. De plus le même tableau, soumis (toujours sans tenir compte de la classification *a priori)* à la classification automatique (agrégation suivant la variance par maximisation de la variance de la partition; cf § 3.8.3) fournit d'abord une dichotomie très nette séparant *I. setosa* des deux autres espèces; puis l'ensemble de celles-ci est partagé en deux, les deux parts coïncidant avec les espèces à quelques erreurs près qu'aucune méthode n'évite et qui semblent dans la nature des choses : *I. versicolor* et *I. virginica* étant morphologiquement contiguës (mais génétiquement distinctes : par le nombre de chromosomes). Ce succès des méthodes inductives (sans information *a priori*) est d'autant plus intéressant que dans bien des cas les classifications préalables, ne sont pas établies avec certitude : il est alors précieux d'établir par le calcul statistique une classification nouvelle issue directement des données descriptives et de celles-ci seulement.

2.3.6. La méthode inductive : En bref, déduire c'est tirer de principes généraux des conséquences particulières; induire c'est s'élever de la connaissance de faits particuliers à celle des notions générales et des liens entre celles-ci. La déduction n'est pas toujours sans imprévu : elle peut, de principes bien connus tirer des conséquences assez surprenantes et assez utiles pour mériter d'être appelées des inventions. Mais l'induction est plus ambitieuse encore : elle n'a pour objet que ce qui nous dépasse. Cependant, une induction n'est confirmée que si les conséquences qu'on en peut déduire sont conformes au réel observable : ainsi induction et déduction ne peuvent être séparées; il faut donc se garder de les opposer. Le statisticien qui dans l'abondance des faits individuels découvre des lignes globales, prétend faire oeuvre inductive : aussi est-il porté à magnifier l'induction. Sans faire fi du raisonnement déductif, Fisher sort toutes ses griffes - qui sont acérées - contre la méthode axiomatique et contre l'exposé déductif de la théorie des probabilités : il met à l'index sans les nommer ces "traités de mathématique" qui ont adopté un traitement formel et abstrait où l'élément d'incertitude est sans effet ("inopérative") parce qu'on évite les applications au monde réel (*Statistical Methods and Scientific Inference* p. 110); et oppose même (*mea culpa* : nous l'avons fait aussi !, et fort radicalement, mais en un autre sens, cf § 1.7) statistique à probabilité (*); au moins si celle-ci est entendue comme une branche autonome des mathématiques (cf § 1.6.2). Claude Bernard au contraire, expérimentateur de génie et théoricien avisé de la méthode hypothético-déductive, se moque de l'engouement des statisticiens pour des moyennes qui ne sont que la confusion des faits réels non leur type idéal (cf § 2.2.1). Reconnaissons qu'aucune méthode ne suffit à conduire systématiquement des données particulières aux idées générales. En critiquant la méthode issue de Fisher, nous affirmons que les outils de calcul disponibles depuis peu permettent d'accomplir rapidement, dans un tout autre esprit, le déploiement ordonné d'un grand ensemble de faits (cf § 3) : nous nous garderons de prétendre accéder infailliblement à l'essence du réel !

Dans la méthode de Fisher, distinguons deux étapes : recueil systématique des données - plan d'expérience (§ 2.3.4); et confrontation de celles-ci à une hypothèse - test probabiliste. Nous rappellerons d'abord la dernière de ces étapes parce qu'elle définit le cadre mathématique de la première.

Fisher refuse le cadre dit bayesien (cf § 1.4.2) des probabilités *a priori* : le modèle universel est pour lui la famille des lois (nous dirons la transition probabiliste; cf § 2.3.3) p_X^0 (dans le cas prédominant de l'analyse de variance, cette famille n'est autre que celle des lois normales sphériques sur R^I). D'après un échantillon $y = \{x^1,\ldots,x^n\}$ il se propose d'estimer la paramètre θ de la loi dont est issu y; ou de discuter une hypothèse nulle telle que $\theta = \theta_0$ (ou, plus généralement: θ est dans telle région O_0 de Θ). Le problème de la validation des hypothèses est pour lui compris dans celui de l'estimation des paramètres que l'hypothèse met en jeu (cf Two new properties of mathematical likelihood; in *Proc. of the Royal Soc., Series A*, T. 144 pp. 285-307, 1934; reproduit dans *Contributions*). Fisher définit un seuil (e.g. 1 %); l'hypothèse nulle est dite rejetée au seuil α si est inférieure à α la probabilité que pour $\theta = \theta_0$ (ou $\theta \in \Theta_0$) la loi p_X^θ conduise à un échantillon y' qui s'écarte plus de l'hypothèse H_0 (c'est-à-dire conduit à une

(*) *Conférence de Fisher à l'Université de l'Etat de Michigan; 1958; cité par Mather et Yates dans leur introduction à l'édition des Collected Papers.*

estimation de θ plus éloignée de θ_0) que ne le fait l'échantillon y, donné. Fisher refuse de définir un risque (ou coût) attaché à l'acceptation ou au rejet de H_0 d'après y; il refuse même de prescrire un seuil universel (α = 1 %; ou α' = 5 %). Son objet premier est de servir le chercheur qui veut découvrir la vérité, non de guider l'homme d'action qui demande une règle de décision : à celui-ci il offre seulement le secours général de la science qui toujours en quelque manière, se montre utile (cf *Statistical Methods and Statistical Inference* pp. 100-104; et *The Design of Experiments* § 12.1). Pour M.G. Kendall (*Biometrika* T. 50 pp. 1-15; 1963; et *Studies*). "C'est en gros le point de vue anglais contre l'américain ... une semblable différence d'attitude est inévitable entre les pays où ce que fait un homme compte plus que ce qu'il pense et ceux où c'est l'inverse". D'une part les Etats-Unis "où sous l'influence de Neyman et de Wald on tend à comprendre la méthode inductive ("inference") dans la théorie de la décision"; de l'autre Fisher selon qui "l'induction n'est pas affaire de décision; et il n'y a en aucun cas pour cela de critère de rentabilité". Il ne nous appartient pas d'arbitrer ce différend : mais nous laissons au lecteur d'imaginer ce que doit être un statisticien français, si en science comme sur la mappemonde, la France s'éloigne des U.S.A. au delà de l'Angleterre !

Conçue par Fisher à propos des recherches agronomiques de Rothamsted (cf § 2.3.4) la méthode du plan d'expérience a été étendue à de multiples disciplines dont Fisher lui-même n'avait pas, sinon par occasions, considéré les données. De la méthode expérimentale issue de Fisher, une image intéressante nous est offerte par Kendall et Stuart (*The advanced theory of statistics* T 3, Griffin, Londres; § 38, pp. 119-123).

D'abord K. & S. opposent expérimentation (experiment) à observation (survey). L'observation porte sur un échantillon extrait d'une population naturelle finie. L'expérimentation se donne pour objet, une population potentielle généralement infinie dont la structure parfaitement régulière sera e.g. celle d'un cube si l'on considère trois variables; et elle produit puis observe un échantillon fini de cas sur lesquels les variables considérées prennent des valeurs spécifiées par tirage aléatoire à partir de la population modèle. Nous avons déjà rencontré cette distinction à propos des données génétiques (§ 2.2.4). L'exemple proposé par K. & S. est l'étude de la dépendance du temps de réaction d'un conducteur de voiture par rapport à la teneur en alcool de son sang. C'est là, selon eux, un problème de régression : exprimer y - le temps de réaction (variable à expliquer) en fonction de x - la teneur en alcool (variable explicative). Comme l'expérimentateur ne peut produire à coup sûr chez un sujet une valeur de x fixée *a priori*, il devra se contenter de fixer dans son plan d'expérience la quantité z d'alcool donnée à chacun. K & S reconnaissent que si on laissait les sujets se servir à boire à leur guise, l'imprégnation alcoolique ainsi réalisée donnerait une plus juste image de ce qui peut se rencontrer dans la pratique (e.g. les grands buveurs, très résistants à l'alcool, prendraient les plus fortes doses). Mais puisqu'il s'agit d'expérimentation, non d'observation, les doses doivent être attribuées aux sujets par tirage au sort.

Cependant le recours au tirage au sort suffit à distinguer le plan d'expérience Fisherien de l'expérimentation préconisée au XIX° siècle par un logicien tel que John Stuart Mill (1806-1873), dont K & S citent cette règle : "Tout phénomène qui varie en quelque manière chaque fois qu'un autre phénomène varie d'une manière particulière, est soit cause, soit effet de ce dernier, à moins qu'il n'y soit relié par quelque fait causal". A cet engrenage de causes supposé universel, mais qui peut se dérober indéfiniment à l'analyse, s'oppose le doute radical de K. Pearson qui dans sa *grammaire* donne pour règle à la *science* de se borner à la description des corrélations (§ 2.2.7). K & S soulignent que Stuart Mill ne conçoit d'étude expérimentale de la relation y = f(x) qu'en interdisant toute variation des autres variables : or d'une part cela est impossible pour un phénomène complexe mettant en jeu de multiples variables (e.g. ci-dessus, l'âge du conducteur, ses habitudes alimentaires, sa profession...), d'autre part l'objet même de l'étude peut requérir la variation simultanée de plusieurs variables (si seule l'association de x_1 à x_2 explique y). Fisher lui-même a écrit de l'étude des causes "one at a time" (cf *The Design of Experiments*, § 37) : "Cette doctrine offre un idéal plus conforme à l'exposé élémentaire d'une théorie physique qu'à la pratique du laboratoire dans quelque branche de la recherche que ce soit".

La méthode du plan d'expérience sera donc d'une part d'éliminer par tirage au sort (affectation aléatoire, *randomisation* disent K & S après Fisher) tout effet systématique des variables qui ne sont pas considérées explicitement dans le plan, mais fournissent un indispensable étalon de variance résiduelle; et d'assigner d'autre part aux variables explicatives mêmes qu'on a choisi d'étudier (x; ou x_1 et x_2; ou z à défaut de x), ainsi qu'à certaines causes d'erreur qu'il importe le plus de neutraliser (comme l'influence des bandes dans le plan en carré latin : cf § 2.3.4) des valeurs judicieusement combinées. On sait que, notamment parce que le coût des expériences en limite le nombre, la combinatoire du plan d'expérience pose aux algébristes d'inépuisables problèmes; de plus K & S rappellent à l'expérimentateur qu'il doit se garder des variables dont le jeu n'est ni réglé par la combinatoire du plan, ni éliminé par affectation aléatoire (éviter e.g. d'examiner les sujets jeunes le matin, et les adultes le soir du fait de leurs obligations professionnelles).

Nous acceptons sans réserve de considérer des cas entre lesquels varient simultanément plusieurs variables. Quant aux démonstrations mathématiques des mérites conjugués de la combinaison systématique et de l'affectation aléatoire, elles ne se lisent pas sans admiration (nous en avons donné un aperçu au § 2.3.4) : on croit saisir une clé miraculeuse. Puis vient le doute : le but de tout plan d'expérience est d'établir une formule de corrélation partielle, entre y, variable à expliquer et e.g. (x_1, x_2) variables explicatives; les autres variables $x_3, \ldots, x_n$ étant éliminées. Or (cf § 2.2.5) la notion de corrélation partielle est très fragile : dans ce problème multidimensionnel il apparaît que l'hypothèse simplificatrice de normalité n'est même pas apte à fournir un modèle qualitatif du réel. De plus, le plan d'expérience ne peut couvrir l'intervalle continu où varie x; il se borne à un très petit nombre de niveaux auxquels sont répétées toutes les mesures : du choix des niveaux dépend le succès de la recherche. Aussi des chercheurs, même parmi ceux fidèles au plan d'expérience, préfèrent-ils aujourd'hui soumettre avant tout à l'analyse des correspondances un ensemble de cas échantillonnant de nombreux niveaux, afin de découvrir rapidement les valeurs charnières des diverses variables.

Efficace dans l'expérimentation agronomique dont Fisher était spécialiste, la méthode du plan d'expérience n'a pas également bien servi toutes les disciplines (auxquelles on l'a étendue d'autant plus volontiers que les traitements et conditions y ont un ensemble fini de modalités; et peuvent donc être des variables qualitatives; cf 2.3.4 *in fine*). Au fond, l'expérimentation à la Fisher, comme celle à la Stuart Mill ne vaut que pour autant que les variables explicitement visées ont été bien choisies, que le lien entre y et x est assez fort : en cela elle n'est pas (comme on l'imagine) un outils suffisant de découverte d'une loi y = f(x); elle précise et confirme seulement la découverte préalable qu'il fallait centrer l'étude sur y et x. Selon nous cette découverte a son origine dans l'observation. Ce qui l'engendre, s'appelle d'un mot simple et grand, *le génie* : libre démarche de l'esprit que ne règle aucun algorithme. Mais à l'examen ordonné d'un ensemble d'observations suggestives, peut aider la statistique : en écologie, en psychologie, en sociologie, partout où les formules y = f(x) ne sont qu'une façon de parler, non une loi déjà découverte, nous préférons généralement l'analyse d'observations judicieusement recueillies sur une base naturelle, à l'expérimentation suivant un plan combinatoire (d'ailleurs souvent inapplicable).

Sir Ronald Fisher est un grand homme ! il croit à la permanence des grands problèmes sur lesquels de siècle en siècle, quelques-uns osent porter une main toujours quelque peu sacrilège. Sans doute doutait-il lui-même qu'il ne fût pas en tout également heureux.