LES CAHIERS DE L'ANALYSE DES DONNÉES

J. P. BENZÉCRI

Histoire et préhistoire de l'analyse des données. Partie II La biométrie

Les cahiers de l'analyse des données, tome 1, n° 2 (1976), p. 101-120

<http://www.numdam.org/item?id=CAD_1976__1_2_101_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° 2 p. 101 120

HISTOIRE ET PRÉHISTOIRE DE L'ANALYSE DES DONNÉES

Partie II La Biométrie

par J. P. Benzécri [1]

2. *Les écoles statistiques anglo-saxonnes* :

2.1. *Pages de chronique* :

Dans ce § nous évoquons une suite de travaux qui de la fin du XIX° siècle au milieu du XX° ont dominé la statistique.

Faire de ces travaux une histoire juste et fidèle requerrait une grande patience dans l'érudition; et pour rendre à chacun son dû, une sainte perspicacité digne du roi Salomon. En de très nombreux mémoires, à propos d'études concrètes, présentées à des publics divers, des idées se font jour dont l'exposé répété se perfectionne graduellement. Les auteurs illustres qui nous ont laissé leur témoignage sur le progrès de la statistique ont plus souvent remémoré les éclairs de leur propre pensée que les reflets qui leur étaient venus d'ailleurs. Reflets d'autant plus fugitifs qu'il n'est pas impossible qu'un statisticien, inspiré par la doctrine d'un psychologue instruit lui-même par un astronome ait cru inventer seul la méthode mathématique utilisée par celui-ci ! Nous tenterons donc seulement de rallumer quelques grandes intuitions que la lourde pratique tabulaire d'il y a vingt ans laisse à peine deviner (**) mais dont l'analyse des données, servie par l'ordinateur nous semble être l'accomplissement : nous saluerons Galton, Pearson, Fisher; en regrettant de ne faire que citer d'aussi distingués savants qu'Edgeworth ou Yule.

Nous parlons d'écoles anglo-saxonnes parce que les travaux les plus influents sont presque tous écrits en langue anglaise par des auteurs britanniques ou américains collaborant parfois avec des collègues d'autres nations; il n'en faudrait pas conclure que ces auteurs sont unis par la discipline ou la fraternité d'une seule école. Ils se combattent sans ménagement ! Parmi les britanniques, R.A. Fisher (cf § 2.3) fut un adversaire irréconciliable de son aîné Karl Pearson; et d'une rive

---

(*) *Suite de l'article paru sous le même titre dans le cahier n° 1 pp. 9-37.*

(1) *Laboratoire de Statistique, Université Pierre et Marie Curie, Paris.*

(**) *De ces intuitions, des notes de Madame Houang Thi sur une conférence de Gower nous ont montré l'intérêt; puis nous sommes sans lassitude remonté aux sources autant qu'il a été possible. Le recueil d'articles historiques parus dans Biometrika et rassemblés par E.S. Pearson et M.G.Kendall y a grandement servi : rappelons que ce recueil intitulé Studies in the History of Statistics and Probability, ed. Griffin, Londres (1970), est cité par nous sous le titre abrégé de Studies.*

à l'autre de l'Atlantique les diverses doctrines de l'analyse factorielle ont garde de former un harmonieux concert (cf § 2.4). Mais les chercheurs de ces écoles rivales bénéficient des travaux les unes des autres; ils ignorent au contraire ce qu'on produit dans d'autres langues en dehors de leurs cercles (*).

A la différence d'un Laplace ou d'un Gauss, ces auteurs ne sont pas de très grands mathématiciens. Intéressés au premier chef par des applications (biométrie, psychométrie, agronomie : qui certes sont la source indispensable de l'inspiration du statisticien), généralement éloignés des travaux de leurs contemporains en algèbre et en analyse, ils utilisent plutôt, parfois avec beaucoup de virtuosité (R.A. Fisher) des techniques déjà classiques. De ce point de vue, les contributions de J. Von Neumann (avec celles de Fréchet ou de Kolmogorov en dehors du monde anglo-saxon) font figure d'exceptions.

Plus que les applications industrielles (contrôle de fabrications) ou économiques, c'est la biométrie (§ 2.2) et la psychométrie (§ 2.4) qui avec la force de vie qui est leur objet, ont inspiré les plus lumineuses intuitions; mais seule la considération des données non-métriques (§ 2.5) a porté la statistique multidimensionnelle à toute sa généralité. Cependant l'ombre de Fisher plane toujours sur nous : plus qu'aucun autre à l'aise dans les espaces multidimensionnels, Sir Ronald y est notre guide; mais en concentrant son attention sur l'étude des lois de forme mathématique exacte (principalement la loi normale, et les lois qui en dérivent) il a proposé aux statisticiens un modèle de la rigueur que nous croyons peu réaliste (§ 2.3). D'où le plan de ce § 2.

2.2. *L'école biométrique anglaise* :

2.2.1. *De Quetelet à Galton* : C'est par l'astronome belge Quetelet (1796-1874) - savant dont l'esprit encyclopédique s'appliqua à la sociologie aussi bien qu'à la géométrie et à la météorologie (**)- que la loi normale s'introduisit dans l'étude des formes vivantes.

On se souvient qu'au nom de Quetelet est associée la doctrine de l'homme moyen qui ne fut pas unanimement acceptée. Il n'est pas de sphère moyenne dont le rayon soit $(R_1 + R_2)/2$, la surface $4\pi(R_1^2 + R_2^2)/2$, le volume $(4/3)\pi(R_1^3 + R_2^3)/2$; "aucune concession n'est possible, ironise Joseph Bertrand dans la préface au Calcul des Probabilités, § IV; nulle sphère n'est difforme. Un homme malheureusement peut l'être, ajoute-t-il, et Monsieur Quetelet en profite; en associant le poids moyen de 20.000 conscrits et leur hauteur moyenne, on fera l'homme type ridiculement gros..." Moins divertissante, mais peut-être plus difficile à éluder, car elle ne met en jeu que des dimensions linéaires, sans y mêler de masse, est cette autre critique : "la hauteur de

(*) *Un cas symptomatique est celui de l'inégalité de Frechet-Darmois-Rao-Cramer, qui borne inférieurement la variance d'une statistique (e.g. un estimateur, cf § 2.3.3) calculée sur un échantillon qui peut être issu d'une loi dépendant d'un paramètre (Fréchet) ou plus généralement de plusieurs (Darmois) : cette inégalité, démontrée d'abord par des auteurs français n'a été employée des statisticiens étrangers qu'après sa redécouverte par des auteurs de langue anglaise, l'un indien - Rao, l'autre suédois - Cramer. Pour un historique précis dû à L. Lebart, cf l'appendice au présent article.*

(**) *On trouvera un panorama de ces travaux dans le recueil Adolphe Quetelet publié à l'Académie Royale de Belgique, Bruxelles, (1974).*

la tête,..., pourra pour l'homme moyen se calculer par deux méthodes : on peut prendre la moyenne des longueurs, ou, pour chaque individu, le rapport de la tête à la hauteur du corps, puis la moyenne de ces rapports" (moyenne qu'on multipliera par la longueur moyenne du corps). "Les résultats sont différents". Nous avons ailleurs opposé cette objection aux hypothèses de normalité : cf T I C n° 5 § 2.2.

Cependant l'originalité de Quetelet n'est pas d'avoir calculé des moyennes en anthropométrie, c'est d'avoir considéré attentivement la dispersion des mesures (e.g. des tailles d'une population d'hommes) et découvert que la loi normale (qu'en tant qu'astronome il connaissait bien) en offrait une description acceptable. Voici en quels termes il présenta sa découverte à son royal correspondant le Duc Régnant de Saxe-Cobourg et Gotha (*) "... supposons qu'on ait employé un millier de statuaires à copier le gladiateur (modèle antique) avec tout le soin imaginable... Je vois sourire Votre Altesse... Je vais peut-être bien l'étonner, en disant que l'expérience est toute faite. Oui vraiment on a mesuré plus d'un millier de copies d'une statue... ces copies étaient même vivantes... J'en viens au fait. On trouve dans le 13° volume du journal médical d'Edimbourg, les résultats de 5738 mesures prises sur les poitrines des soldats des divers régiments écossais... La différence que la nature met entre les tailles des hommes n'est pas plus grande que celle que produirait l'inexpérience dans les mesures prises sur un même homme ayant une attitude plus ou moins courbée... Tout se passe comme s'il existait un homme type, ... : chaque peuple présente sa moyenne et les différents écarts de cette moyenne en nombres calculables a priori (d'après la loi normale). Cette moyenne varie d'un peuple à l'autre". Quetelet sait l'importance d'un échantillonnage naturel (cf § 2.2.4; 2.3.6; 3.7.1). "Quand on veut une vérification de la loi... il ne faut point choisir, il faut prendre tous les hommes d'une nation tels qu'ils sont". Et il a vu avant Pearson que certaines causes pouvaient détruire la symétrie de la distribution normale par des écarts de grande amplitude soit au delà soit en deçà du centre de la distribution.

L'étape suivante, l'étude conjointe de la variation de deux mesures (e.g. de la taille du père avec celle de son fils; ou de la longueur du bras avec celle de la jambe d'un même homme) fait la gloire de l'anglais Galton. Charles Darwin, dont Galton était le cousin, avait affirmé (cf § 2.2.4) que les espèces évoluaient de génération en génération parce que les individus les mieux adaptés au milieu ayant la plus longue vie et la plus abondante progéniture, leur type tendait à prédominer. C'était poser au statisticien le problème de la comparaison des enfants aux parents - car sans une ressemblance étroite de ceux-là à ceux-ci, les caractères distinctifs des mieux adaptés d'une génération n'auraient pas passé à la suivante -; et aussi celui de la diversité des formes au sein d'une génération - car sans une telle diversité, il n'y aurait pas d'individus supérieurement adaptés dont au travers des hasards de l'existence le type dut s'imposer -. La théorie de Darwin a inspiré de grands enthousiasmes, et même des passions durables qui soutinrent l'extraordinaire labeur de l'école biométrique issue de F. Galton. Il ne nous appartient pas d'apprécier ce que la biologie a acquis par ce labeur (nous oserons seulement en parler en passant, pour tirer des leçons utiles au statisticien : cf § 2.2.4) : mais nous pouvons affirmer que dans leur ardente quête des chiffres,

---

(*) *Lettres à S.A.R. le Duc Régnant de Saxe-Cobourg et Gotha; sur la Théorie des Probabilités appliquée aux sciences morales et politiques; Bruxelles (1846)*

quelques hommes inégalement instruits en mathématiques, et dont les plus géomètres ne se peuvent comparer à un Laplace ou à un Gauss, ont au fil des années découvert des lois, posé et parfois résolu des problèmes que Poincaré leur contemporain eût dominé en quelques mois; mais qu'aucun grand mathématicien n'eut le mérite de rencontrer : c'était avoir bien mérité de la science.

2.2.2. Régression et corrélation découvertes en biologie : Quand Galton entreprit de comparer les caractères des enfants à ceux des parents (*), les documents antropométriques manquaient totalement pour cela : il commença donc par l'étude des petits pois; acquit des graines sélectionnées, les mesura, les sema et mesura les graines de sa récolte. Les résultats des mesures peuvent être consignés dans un tableau rectangulaire; comme notre propos est de faire l'histoire de l'analyse des données, on nous pardonnera de décrire ici sans épargner les détails un tableau de correspondance qui devait jouer un grand rôle dans les recherches de Galton puis de Pearson, avant de devenir notre héritage (cf §§ 3.4.5 & 3.5.2.)! Soit J la suite des valeurs entières j que peut prendre le diamètre (compté en centième de pouce : environ 0,27 mm) d'une graine de semence; et de même I pour le diamètre i d'un pois de la récolte : on forme un tableau I x J et inscrit à l'intersection de la ligne i et de la colonne j le nombre $k(i,j)$ des pois de la récolte dont le diamètre est $i((i \pm 0,5)10^{-2}$ pouce) et qui sont issus d'un pois de diamètre $j((j \pm 0,5)10^{-2}$ pouce). Pour chaque colonne (c'est-à-dire pour les pois issus d'une semence de diamètre j donné) Galton calcula la moyenne et la variance de i (i.e. du diamètre des pois récoltés). Il trouve une moyenne $\bar{i}(j)$ approximativement fonction linéaire de j ce que nous écrivons en une formule (où $\bar{i}$ et $\bar{j}$ désignent respectivement le diamètre moyen de l'ensemble des pois récoltés et des semences d'où ils sont issus; et r un coefficient sur lequel on reviendra) : $\bar{i}(j) - \bar{i} = r(j - \bar{j})$. Et de plus "was certainly astonished to find the variability of the produce of the little seeds to be equal to that of the big ones..." : la variance de la colonne ne dépendait pas de j.

Aujourd'hui, on enseigne classiquement que si x et y sont deux variables aléatoires de variance σ^2 (variables que nous supposerons de moyenne nulle pour simplifier l'écriture) dont le coefficient de corrélation est r, la loi conjointe de x et y est, sous l'hypothèse de normalité, donnée par la formule :

$$(2\pi\sigma^2 (1 - r^2)^{1/2})^{-1} \exp(-(x^2 - 2rxy + y^2)/(2\sigma^2(1 - r^2))) \, dx \, dy;$$

fixons x: la loi conditionnelle de y (loi de y dans la tranche x, x + Δx) n'est autre que :

$$(2\pi\sigma^2 (1 - r^2))^{-1/2} \exp(-(y - rx)^2/(2\sigma^2(1 - r^2))) \, dy;$$

(cette formule résulte simplement de ce que $x^2 - 2rxy + y^2 = (y - rx)^2 + (1 - r^2)x^2$; donc pour x constant, la densité est proportionnelle à $\exp(-(y - rx)^2/(2\sigma^2(1 - r^2)))$.) et de même si l'on fixe x la loi conditionnelle de y est :

(*) *Nous suivrons ici K. Pearson : Notes on the History of Correlation; Biometrika T. 13, pp. 25-45 (1920) reproduit dans Studies.*

$$(2\pi\sigma^2\ (1 - r^2))^{-1/2}\ \exp(-(x - ry)^2/(2\sigma^2\ (1 - r^2)))\ dx;$$

Ainsi, x étant fixé, la moyenne de y est rx, et la variance de y est $\sigma^2(1 - r^2)$ (indépendante de x), et de même pour y fixé, x a moyenne ry et variance $\sigma^2(1 - r^2)$. La variance conditionnelle (variance sur une tranche) est le produit par $(1 - r^2)$ de la variance globale σ^2 (variance sur l'ensemble des individus). En dix ans, sans s'aventurer dans l'analyse mathématique, par la seule élaboration des données qu'il rassemblait, Galton allait retrouver ce petit formulaire (et même un peu plus : nous avons fait l'hypothèse restrictive que x et y ont même variance σ^2; tel n'est pas toujours le cas). "That Galton should have evolved all this from his observations - s'exclame Pearson - is to my mind one of the most note - worthy scientific discoveries arising from pure analysis of observations". Dix ans encore, et Pearson lui-même possèderait le formulaire analogue pour la loi normale multidimensionnelle (cf § 2.2.5).

Ayant trouvé pour les diamètres des pois r = 0,33, Galton souligne d'abord que l'écart $(\bar{i}(j) - \bar{i})$ des fils (les pois de sa récolte) à la moyenne, n'est que le tiers de l'écart $j - \bar{j}$ des pères (les semences) à la moyenne : il y a donc eu d'une génération à la suivante retour vers la moyenne, vers le type. En 1877, devant la Royal Institution of Great Britain, Galton parle de reversion, on parlera ensuite de régression (dans un sens élargi : pour toute formule approchée exprimant une variable aléatoire à expliquer, en fonction d'autres variables aléatoires dites explicatives; comme ici le diamètre des fils assimilé à $\bar{i}(j)$ est exprimé en fonction linéaire du diamètre j du père) : la lettre r en est restée pour désigner le coefficient de corrélation. De plus Galton a compris que la variance sur une tranche doit être le produit de la variance d'ensemble σ^2 par $(1 - r^2)$: car la variance d'ensemble σ^2 est stable de génération en génération; or la variance des fils est la somme de la variance héritée des pères, réduite dans la proportion r^2, et de la variance sur une tranche (variance pour les descendants de pères dont la taille est fixée) : cette dernière variance est donc $\sigma^2 - r^2\sigma^2 = (1 - r^2)\sigma^2$. On voit que Galton sait que la variance d'une somme de termes (supposés non corrélés) est somme de la variance des deux termes. C'est tout pour 1877.

Cependant en encourageant les sujets par des primes (il invite les famille à accepter qu'on les mesure...), Galton rassemble des données anthropométriques : il est à même de constituer des tableaux de correspondance I x J (cf supra) où il ne s'agit plus de récolte et de semence, mais (comme nous le suggérions déjà par les mots de fils et de père) d'enfants et de parents. Les nombres inscrits dans les cases du tableau I x J sont (à un coefficient près; et si l'on fait abstractions des fluctuations d'échantillonnage) les valeurs de la fonction densité $\exp(-(x^2 - 2\ rxy + y^2)/(2\sigma^2(1 - r^2)))$ que nous avons rappelée plus haut. Empiriquement, Galton trace les courbes de niveau de cette fonction : il découvre les ellipses homothétiques et concentriques !

(les courbes d'équation $x^2 - 2\,rxy + y^2 = Cte$); de plus, la ligne de régression (Galton put marquer en gris dans chaque colonne j du tableau la case la plus lourde, qui est aussi celle de la valeur moyenne $\bar{i}(j)$; puis marquer exactement le point moyen; ces points s'alignent sur une droite de pente r) n'est autre que le diamètre conjugué par rapport à ces ellipses de la direction des colonnes (cf fig. 2-1); enfin chaque colonne ainsi que chaque ligne a le profil d'une loi normale de variance constante.

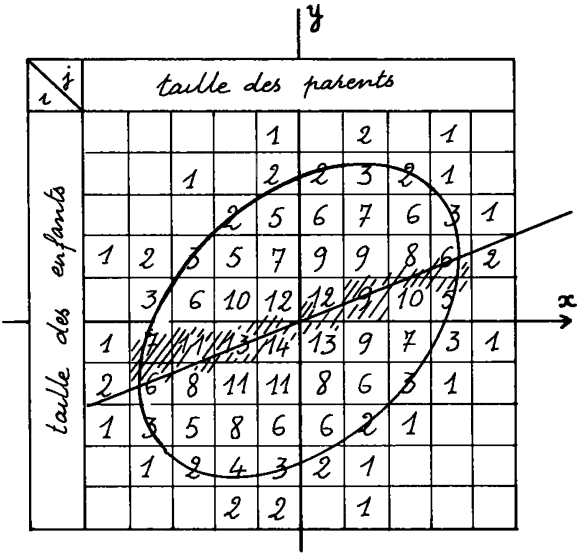

Figure 2 1 Du tableau de correspondance à la fonction de densité de la loi normale bidimensionnelle, d'après Galton, on a modifié le tableau de Galton (reproduit dans Pearson, Biometrika, T 13, 1920, et Studies p 196) afin de suivre exactement les notations de notre exposé et l'hypothèse simplificatrice de variance égale en x et en y. Le lecteur remarquera qu'il s'en faut de beaucoup que le dessin d'ellipses d'égale densité s'impose au regard

2.2.3. Formules mathématiques et ordre de la nature :

De toutes ses remarques Galton fit part à un mathématicien de Cambridge J.D. Hamilton Dickson, qui en déduisit sans retard la formule de la loi normale bidimensionnelle et démontra avec précision les résultats que Galton n'avait atteints que par une interpolation délicate (with tender caution, dit-il lui-même; notamment en prenant les moyennes quatre à quatre des cases du tableau de correspondance, pour régulariser les données de ses comptages). Et Galton de s'émerveiller (dans son Adresse à la Section d'Anthropologie de l'Association Britannique à Aberdeen; 1885)"...I never felt such a glow of loyalty and respect towards the sovereignty and magnificient way of mathematical analysis..." Qui a été comme nous, élevé sous le sceptre de N. Bourbaki, s'étonnera plutôt que les mathématiciens n'aient pas découvert par eux-mêmes toutes les propriétés de la loi normale multidimensionnelle; que Galton n'ait pas été devancé. En effet nous avons vu (cf § 1.5.3') que la formule de la loi normale multidimensionnelle est explicitement chez Laplace; et dans l'étude même des corrélations, Galton et Pearson eurent un prédécesseur : Auguste Bravais.

Physicien, astronome, géologue, géodésiste, probabiliste, ... Auguste Bravais (1811-1863) est l'auteur d'un travail "Sur les probabilités des erreurs de position d'un point" publié en 1846 (dans le recueil des Mémoires présentés par divers savants à l'Académie royale des Sciences de l'Institut de France; T IX pp 256-332). De ce travail communément donné comme référence princeps à l'étude de la loi normale plane et spatiale, le texte est peu accessible : aussi en donnerons-nous une brève analyse illustrée de quelques citations. Voici comment dans un Résumé Bravais définit le programme de ses recherches:"...l'illustre Laplace... s'est borné à l'appréciation de la possibilité des erreurs de la mesure finale d'une longueur, c'est-à-dire des erreurs qui peuvent exister dans la situation d'un point assujetti à rester sur une droite donnée... j'ai d'abord assujetti le point cherché à varier dans un plan donné, et j'ai examiné ensuite le cas général de l'espace".

A la vérité, s'il avait saisi dans le détail l'exposé de Laplace, Bravais aurait dû borner ses prétentions ! car Laplace traite explicitement de la loi de plusieurs mesures (cf § 1.5.3'). Mais considérons qu'après 150 ans de réflexions sur les probabilités et la statique, il est encore difficile d'embrasser les trésors que Laplace enferme dans ses mystérieuses démonstrations; et pardonnons à Bravais dont, quant à l'étude des corrélations au moins, le travail est original.

Bravais part des formules obtenues par Laplace d'après le critère des moindres carrés. Il suppose que sont normales centrées (Laplace quant à lui n'a pas besoin de l'hypothèse de normalité peu réaliste à ce niveau (*); il démontre par le théorème central limite que les estimations des grandeurs inconnues sont distribuées normalement, quelle que soit la loi des erreurs de mesure) les erreurs, notées m, n, p,..., commises sur les mesures primaires (d'angles, de longueurs; voire de temps); les erreurs (x,y) sur les coordonnées d'un point du plan, ou celles (x,y,z) sur les coordonnées d'un point de l'espace sont alors de la forme :

$$x = A\,m + B\,n + C\,p + \ldots ;$$
$$y = A'm + b'n + C'p + \ldots ;$$
$$z = A''m + B''n + C''p + \ldots ;$$

et la loi de x est $(h_x/\pi)^{1/2} \exp(-h_x x^2)dx$, où h_x est donné par :

$$1/h_x = (A^2/h_m) + (B^2/h_n) + (C^2/h_p) + \ldots = \Sigma(A^2/h_m)$$

formule où h_m, h_n, h_p ... est le paramètre de la distribution de m, n, p; paramètre que Bravais appelle module des erreurs possibles, et que nous écrivons $(1/2)\sigma_m^2$... (Notons ici que sans parler explicitement de variance Bravais se demande pourquoi la crainte mathématique du carré de l'erreur ne pourrait pas aussi bien que l'espérance mathématique du module de celle-ci servir à mesurer la précision de l'observation ...). On a de même :

$$1/h_y = (A'^2/h_m) + (B'^2/h_n) + (C'^2/h_p) + \ldots$$

(*) *J. Bertrand lui-même n'a pas vu cet avantage des méthodes de Laplace, car il écrit : "Les études faites sur cette question (des erreurs de situation d'un point), particulièrement par Bravais dans un Mémoire remarqué, supposent une confiance absolue dans la loi proposée par Gauss sur la probabilité des erreurs élémentaires... Les formules déduites de cette hypothèse sont confirmées par les faits connus". Voila qui nous confirme que tandis que le mémoire de Bravais fut remarqué les démonstrations de Laplace ne furent guère lues : elles seules expliquent cependant, pourquoi "les formules sont confirmées par les faits".*

Mais, remarque Bravais,"la coexistence des mêmes variables m, n, p... dans les équations simultanées en x et y, amène une corrélation telle, que les modules h_x, h_y, cessent de représenter la possibilité des valeurs simultanées de (x,y) sous le vrai point de vue de la question". Voilà innocemment prononcé ce mot de corrélation dont la carrière en statistique devait être brillante ! Cependant Bravais poursuit en cherchant la loi conjointe de (x,y). Il démontre d'abord qu'en effectuant une substitution linéaire sur les m, n, p, on obtient pour densité en les nouvelles variables m', n', p' l'exponentielle d'une forme quadratique; puis que cette forme subsiste quand par intégration partielle on élimine une ou plusieurs des variables, afin d'avoir la loi conjointe des variables restantes. D'où la forme de la loi de (x, y) :

$$(K/\pi)\ e^{-(ax^2 + 2cxy + by^2)}\ dx\ dy = \omega(x,y)\ dx\ xy$$

Reste à déterminer a, b, c, K, en fonction des A, B, C,..., A', B', C'. Compte tenu de ce que l'intégrale de $\omega(x,y)$ est 1, et de ce qu'il connaît la distribution de x et y, Bravais obtient sans peine trois relations :

$$a = K^2 \Sigma(A^2/h_m) = K^2/h_x;$$

$$b = K^2 \Sigma(A'^2/h_m) = K^2/h_y;$$

$$K^2 = ab - c^2;$$

pour calculer c, Bravais a une très heureuse inspiration; il fait dans le plan des (x,y) une rotation des axes; ainsi d'une part il découvre les axes principaux des ellipses d'égale densité (c'était faire la première analyse factorielle, cf infra §§ 2.2.5 & 2.4); d'autre part comme le paramètre h, ($h = 1/2\sigma^2$) d'une combinaison quelconque $x \cos\alpha + y \sin\alpha$ n'est autre que $\Sigma(A \cos\alpha + A' \sin\alpha)^2/h_m$, Bravais peut achever de déterminer $\omega(x,y)$; et brille dans ses formules, au côté de $\Sigma(A^2/h_m)$ et de $\Sigma(A'^2/h_m)$ qui sont (à un facteur 2 près) des variances, la somme $\Sigma(AA'/h_m)$ qui est une covariance et dont la nullité (Bravais l'a vu) est la condition d'indépendance entre les deux coordonnées.

Nous ne détaillerons pas les divers problèmes de probabilité que Bravais résoud en appliquant ses formules; mais nous esquisserons ses résultats en dimension 3. Bravais atteint la formule de la loi $\omega(x,y,z)$; propose ici encore de passer aux axes principaux; et ayant trouvé une formule analogue à $K^2 = ab - c^2$ du cas bidimensionnel, Bravais s'enhardit à calculer qu'en dimension quatre également, le coefficient placé devant l'exponentielle est la racine carrée du discriminant de la forme quadratique placée en exponentielle; mais reconnaît que "la démonstration générale de cette loi de formation lui est inconnue".

Pourtant, K. Pearson refuse que Bravais ait découvert la corrélation; en ce sens que pour celui-ci, comme pour les autres spécialistes de la théorie des erreurs, x et y (ou x,y et z; etc) ne sont que des valeurs estimées d'après des mesures de base (distances ou angles) qui, elles, sont des variables aléatoires normales (centrées sur la vraie valeur s'il n'y a pas d'erreur systématique) deux à deux indépendantes : "que les quantités directement mesurées puissent être elles-mêmes corrélées, ne leur est pas venu à l'esprit..." Ici, (cf Seal in Biometrika T 54, pp 1-24, 1967; et Studies).

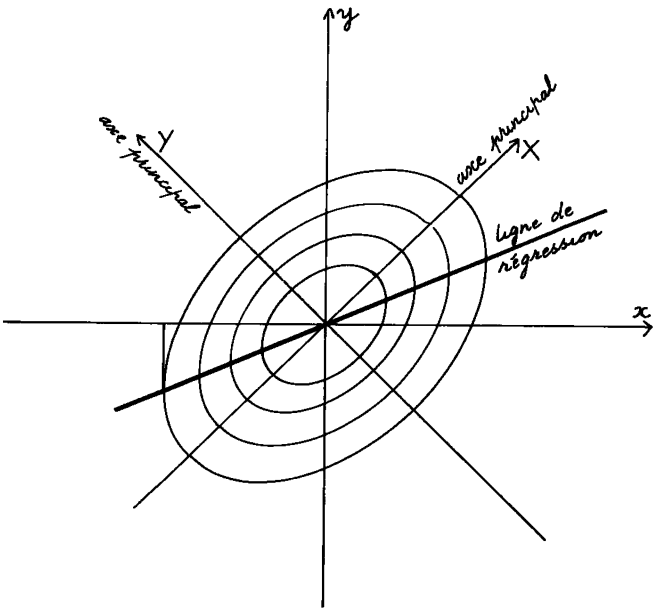

Figure 2.2. Etude géométrique de la loi normale bidimensionnelle ; ellipses d'égale densité et axes principaux déjà vus par Bravais; la ligne de régression apparait aussi incidemment dans son travail ; mais sans son interprétation explicite

K. Pearson manque de citer le Hollandais Schols qui dès 1875 note que les composantes verticales et horizontales de l'erreur de tir d'un canon ne sont pas indépendantes. De plus les mesures corrélées rentrent dans le cadre des mesures indépendantes, car, derrière des quantités corrélées x^1, x^2,...; on en imaginera toujours d'autres u^i indépendantes entre elles, mais inaccessibles aux mesures; et on postulera un modèle linéaire : $x^h = \Sigma A_i^h u^i$, analogue à celui qui dans la théorie des erreurs relie les coordonnées cherchées x^h, aux observations u^i. Une même vue mathématique peut tout comprendre. Mais encore faut-il y parvenir

Cette idée "que les quantités directement mesurées puissent être elle-mêmes corrélées", si fondamentale pour l'analyse des données, était assurément plus cachée que nous ne pouvons l'imaginer, car Galton lui-même n'acheva de l'acquérir qu'en 1888. Jusque là Galton n'avait considéré que la loi conjointe d'une mesure x prise sur le père (ou de la moyenne des parents) et de cette même mesure y prise sur l'enfant : en 1888, il découvre que l'appareil mathématique qu'il a patiemment monté peut servir à étudier les variations conjointes de deux organes sur une population d'individus (e.g. x largeur de la main; y largeur du pied etc). Ce que les biologistes appellent corrélation de structure (en 1888 on écrit aussi co-relation) est désormais mesurable par le paramètre r d'une loi normale bidimensionnelle.Dans Natural Inheritence (1889) Galton propose un formidable programme de recherche biométrique : étudier statistiquement la variabilité et la plasticité des formes vivantes, afin de confirmer mathématiquement le mécanisme de l'évolution dessiné par Darwin.

2.2.4. *K. Pearson et R. Weldon* : la lecture de Natural Inheritence suscita à F. Galton deux disciples dont l'ardeur était à la mesure du travail proposé : K. Pearson (1857-1936) et R. Weldon (1860-1906). Travailleur inlassable, aussi curieux de philosophie ou de droit romain que de mathématique et de physique, se passionnant pour la transformation de la société (on a pu supposer, cf Haldane in Biometrika T 55 1957 ou Studies p. 428, que c'est en hommage à Marx qu'il orthographiait Karl son prénom de Charles; encore que sa doctrine ne fût pas rigoureusement marxiste), K. Pearson fut dès 1884 professeur de mathématiques appliquées à University College de Londres et eut bientôt la liberté de se consacrer largement à la philosophie naturelle. Dessinateur de talent, non moins cultivé que Pearson, aussi adonné que lui au labeur, R. Weldon obtint en 1890 une chaire de zoologie à University College. De leur rencontre dans ce College (analogue par sa fonction à ce que nous appellerions en France une grande école; mais d'un autre esprit) jusqu'à la mort prématurée de Weldon (emporté par une pneumonie en 1906) Pearson et Weldon collaborèrent fièvreusement : celui-ci s'appliquant à fournir par la biométrie des preuves expérimentales rigoureuses de la doctrine de l'évolution; celui-là perfectionnant l'outil statistique pour élaborer de telles données et en critiquer l'interprétation (cf §§ 2.2.5 & 2.2.6).

Il ne nous revient pas de décrire les travaux de Weldon : mais il importe à l'analyse des données d'en citer des exemples pour en méditer les leçons. Dès 1892 (cf K. Pearson, in Biometrika T 6 1906 pp 1-52 et Studies, pour une biographie de Weldon et l'histoire de ses travaux) Weldon publie pour cinq races locales de crevettes les coefficients de corrélation entre les dimensions de 22 paires d'organes, déterminés d'après 1000 individus de Plymouth, 500 de Roscoff, etc. Il en conclut que ces coefficients r ne varient pas de race à race : conclusion qu'après Pearson le statisticien contemporain regardera comme illusoire. Mais ce sont justement les données de Weldon qui requirent de créer la théorie des erreurs d'échantillonnage sur r(*) : en 1892, le coefficient de corrélation est une nouveauté, et Weldon l'appelle encore "fonction de Galton". Aujourd'hui (cf §§ 3.6.1 & 3.7.2) il nous plairait de soumettre à l'analyse des correspondances l'ensemble des données de Weldon : nous déterminerions les axes factoriels de la population totale et ceux propres à chacune des races locales; critiquerions leur stabilité sans hypothèse de normalité en répétant l'analyse sur des sous-échantillons successifs, comparerions les races entre elles... Weldon, quant à lui n'avait pas d'ordinateur (l'outil rêvé était alors la calculatrice mécanique Brunswiga) : et après avoir recueilli (élevé parfois) puis mesuré les crustacés, il devait encore calculer. Ainsi débutait l'étude de la variabilité des formes vivantes pour une autre espèce que l'homme : sur le modèle de l'anthropométrie (dont la pratique remonte au moins à Quetelet) se fondait la biométrie.

Après les crevettes, les crabes : comparaison entre la race de Plymouth et celle de Naples, d'une même espèce : ici encore, la matrice de corrélation est stable de race à race. Et finalement, après la variabilité synchronique, la diachronie, l'évolution, la sélection. De 1892 à 1896, les formes moyennes des crabes du Plymouth Sound se modifient : Weldon entrevoit une corrélation entre la largeur frontale (liée à l'appareil respiratoire) et le taux de survie dans des eaux de plus en plus polluées. Il organise un élevage de crabes dans un milieu dont le contact eût suffit à rebuter tout autre que lui et finalement établit avec certitude que les générations de survivants ont un rapport (largeur

---

(*) *Weldon lui-même, en 1894, recourut à une étude empirique des fluctuations, en faisant recenser par un secrétaire les résultats de dizaines de milliers de lancers de dés ! un extrait de ces données est analysé par R.A. Fisher dans son très classique Statistical Methods for Research Workers.*

frontale/Largeur totale de la carapace) plus faible en eau turbide qu'en eau claire. Sans adopter le darwinisme (la plasticité des espèces est un fait : mais d'une part l'explication qu'en a donnée Darwin n'est plus admise, cf infra; d'autre part la genèse des formes vraiment nouvelles requiert plus que ce seul fait) on peut s'émerveiller de l'exploit expérimental que cette doctrine avait suscité !

Aux conceptions génétiques de Pearson et de Weldon, le statisticien de 1975 doit ici s'arrêter, parce que de nos jours la génétique n'est pas une science achevée et que dans l'interprétation de grands ensembles de données nouvelles peuvent se reproduire des malentendus analogues à ceux du début du XX° siècle.

En calculant chez l'animal comme chez l'homme le coefficient de corrélation entre une mesure prise chez un parent et un enfant, ou chez deux frères, Galton puis Pearson ont très souvent trouvé des valeurs de r autour de 0,5 : ils sont parvenus à la conviction qu'à des variations de faible amplitude près, une part constante de la variance de la génération des enfants est héritée des générations antérieures successives. Telle est en substance la Law of Ancestral Heredity dont l'élaboration précise requit des calculs de corrélations partielles dans le cadre du modèle normal (§ 2.2.5); calculs qui préfigurent parfois l'analyse factorielle (§ 2.4). Pearson insiste sur le fait que cette loi de l'hérédité ne dit rien de la différence entre moyenne des caractères pour deux générations successives d'une même population : elle spécifie seulement une relation statistique (corrélation) entre d'une part l'écart d'un enfant par rapport à la moyenne de sa génération et d'autre part les écarts de ses parents et ancêtres antérieurs par rapport à la moyenne de la leur.

Quand fut énoncée la Loi de l'Hérédité Ancestrale, la génétique allait connaître un événement extraordinaire : en 1900 les lois de Mendel - publiées dès 1865; mais non remarquées alors - sont redécouvertes par De Vries, Tschermak, Correns. Pour comprendre les controverses qui opposèrent Galton et Pearson aux mendéliens, il faut tenter de ranimer la pensée génétique d'avant 1900 : un petit volume de Félix Le Dantec (biologiste préoccupé de philosophie; cité au § 1.6 à propos d'un paradoxe des probabilités) - Lamarckiens et Darwiniens - publié justement en 1899 (chez F. Alcan) nous y aidera.

L'adaptation des espèces animales et végétales à leur milieu a toujours été remarquée des philosophes : mais l'idée que de cette harmonie le milieu lui-même en modifiant les espèces, pût être la cause instrumentale est associée aux noms de Lamarck (1784-1829) et de Darwin (1809-1882). Il est d'usage depuis longtemps d'opposer ces deux noms : à parcourir l'histoire on voit s'estomper cette opposition : Darwin a seulement eu quelques idées de plus que Lamarck et a pu ainsi restreindre la portée de celles de son devancier (dont il n'honora nullement la mémoire) sans les rejeter pourtant absolument; mais les générations successives de néo-darwiniens ont amalgamé au darwinisme toutes les idées intervenues depuis,faisant de Darwin le patron universel; et de Lamarck son négatif. L'homme sait depuis des millénaires disposer d'animaux et de plantes meilleurs par deux moyens : l'élevage (terrain et engrais pour les plantes; nourriture et exercices pour les animaux; et aussi les hommes) et la sélection (choix des meilleurs; rejet, voire destruction des autres). Il était naturel de penser que la Nature agissait de même pour produire l'évolution des vivants. Lamarck pensa à l'analogie de l'élevage : il postula que les caractères (allongement de telle partie; réduction de telle autre) acquis par un individu à l'épreuve de la vie, passaient à sa descendance; en sorte que de génération en génération l'évolution pouvait par degrés aboutir très loin de

sa source. Darwin pensa à la sélection : la nature fait périr tôt les moins adaptés, elle donne aux mieux adaptés une large descendance, ainsi se déplace le centre de gravité de l'espèce. "J'ai donné, écrit-il (nous citons d'après Le Dantec dans la traduction de Barbier), le nom de sélection naturelle ou persistance du plus apte à la conservation des différences et des variations individuelles favorables et à l'élimination des variations nuisibles..." Mais il a écrit aussi : "Ces modifications ont été effectuées principalement par la sélection naturelle de nombreuses variations légères et avantageuses; puis les effets héréditaires de l'usage et du défaut d'usage des parties ont apporté un puissant concours à cette sélection". Ce qui laisse place à des idées qu'un contemporain croirait n'appartenir qu'à Lamark.

En des termes qui nous sont familiers, on peut dire que l'organisation complexe du vivant relève de l'information : cette information anime le corps dans son ensemble (Aristote dit que l'âme est la forme du corps); n'y est-elle pas aussi matériellement inscrite en de petits corpuscules, particulièrement dans la semence ? Déjà Buffon avait produit la théorie des molécules organiques : théorie fabuleuse, mais où le spécialiste de biologie moléculaire s'émerveille de trouver préfigurés le codage et la réplication. Darwin postule l'existence de gemmules; "chaque cellule de l'organisme produit un grand nombre de gemmules qui toutes représenteraient exactement la cellule où elles sont nées telle qu'elle était au moment de leur naissance, et qui toutes, voyageant ensuite à travers l'organisme, auraient la vertu spéciale de donner à toute cellule neutre dans laquelle elles pénétreront les caractères de la cellule d'où elles proviennent" (Le Dantec; op. laud. p. 51). Voilà un support pour l'hérédité, et l'hérédité des caractères acquis puisque les gemmules sont une image actuelle des cellules... Mais Galton par des expériences de transfusion que le titre de son mémoire suffit à suggérer : "Experiments in Pangenesis by breeding from rabbits of a pure variety, into whose circulation blood taken from other varieties had previously been largely transfused" (Proc. Roy. Soc., 1871) estime avoir réfuté la circulation des gemmules. Avec elles disparait la possibilité d'un flux des caractères acquis vers la semence des vivants. Mais quand écrit Le Dantec (1899) De Vries reprend "les gemmules (non circulantes) en les mettant au courant des découvertes les plus récentes de l'histologie". La théorie contemporaine de l'hérédité chromosomique et mendélienne est à l'horizon !

Le modèle génétique de Mendel, sous sa forme la plus simple suppose que tout individu possède de chaque caractère élémentaire (associé à ce qu'on nomme aujourd'hui locus chromosomique) deux modalités (appelées allèles ou gènes allélomorphes de ce locus) portées par deux chromosomes homologues, chacun reçu d'un parent (qui lui-même possède deux allèles, ses chromosomes se distribuant par paires homologues; mais n'en transmet qu'un à chaque enfant). Ainsi contrairement à ce qu'on pensait implicitement jusqu'alors, l'enfant n'est pas comme une moyenne entre ses parents réalisée par fusion; c'est un mélange qui a reçu intact de chacun de ses parents la moitié de ses composants géniques. L'exemple type de Mendel est celui où pour un locus il n'y a que deux allèles possibles; l'un dominant D, l'autre récessif r : un sujet possédant deux fois D (DD) ou une fois D (Dr) a l'apparence liée à D; tandis que l'apparence liée à r n'appartient qu'aux sujets (rr).

Voici de toutes autres lois de l'hérédité que celle des biométriciens. Certes Mendel n'aurait pu aboutir à ces lois sans d'amples statistiques, et leur énoncé même rappelle une formule probabiliste. Mais en opérant dans des conditions expérimentales très particulières (au début deux races pures de pois; puis entre générations des fécondations contrôlées : une plante destinée à porter fruit étant privée de ses étamines à temps

pour qu'elle ne puisse être fécondée que par le pollen qu'on a choisi d'y porter), Mendel a créé une population (une suite de générations) dont la structure est très artificielle; et c'est seulement ainsi que les lois sont apparues comme évidentes (du moins au génie de Mendel).

Pour Pearson la découverte de Mendel a peut-être une portée universelle; mais l'affirmer est une théorie qu'il juge prématurée (*). D'ailleurs l'unanimité est loin d'être faite, alors, en faveur des mendéliens parmi les biologistes. On le comprend si l'on sait que les expériences à la Mendel sont très difficiles à contrôler; que le modèle Dr n'est pas le modèle unique (il existe des locus auxquels correspondent plus de deux gènes allélomorphes); et qu'à la différence des qualités considérées par Mendel (e.g. graine lisse ou graine profondément ridée; cotylédone vert ou cotylédone jaune) beaucoup d'aspects macroscopiques d'un vivant (e.g. la couleur de la peau; et maint caractère variant continuement) ne peuvent s'expliquer par un seul locus (on doit leur attribuer des modèles complexes à plusieurs locus)... Pearson et Weldon (**) tentent pourtant de confronter avec un modèle mendélien généralisé (affectant plusieurs locus à un caractère continu), leur loi de l'hérédité ancestrale : l'accord n'est qu'approximatif (r = 0,42 entre frères; r = 0,33 entre père et fils). Pearson conclut qu'il n'y a pas d'opposition essentielle entre les lois de corrélation de la biométrie et les mécanismes mendéliens; mais que la conception qu'on se fait de ceux-ci doit être assouplie. Quand meurt Weldon, en 1906, l'opposition est déjà dure entre Bateson, chef de file des mendéliens anglais, et Pearson. Celui-ci n'est pas à même de poursuivre seul ces délicates recherches : E.S. Pearson (op. laud. p. 241), fils de Karl,pense que si Weldon eût vécu, l'harmonie se serait faite plus tôt entre la biométrie des populations naturelles et les modèles génétiques aussi carrés que le jeu de dé ! découverts par le génie expérimentateur de Mendel sur des populations absolument planifiées. Il était réservé à Fisher de conjuguer rationnellement ces doctrines (cf § 2.3.1).

Assurément, l'expérimentation mendélienne s'oppose au principe de la biométrie (principe fondamental en analyse des données, cf. e.g. § 3.7.1) selon lequel la composition de la population (ou de l'échantillon qui la représente) est elle-même un phénomène naturel; et de plus, digne d'être étudié. Il ne faut évidemment pas en conclure que les faits découverts par Mendel sont faux ou inintéressants. En général les échantillons non-naturels brouillent tout et ne découvrent rien; mais c'est aussi par des constructions expérimentales très éloignées de ce qu'offre la nature laissée à elle-même, non-contrainte (***), qu'ont été découvertes les lois de la physique, par exemple celles du courant électrique. Nous reviendrons au § 2.3.6, à propos de l'exposé que donne Kendall de la méthodologie issue de Fisher, sur cette si importante distinction entre observation (de populations naturelles) et expérimentation (sur des échantillons artificiels; voire dans des conditions de contrainte).

(*) *E.S. Pearson, Biometrika T 28, p. 219 (1936) souligne qu'il est conforme à la philosophie de son père (auteur de The Grammar of Science, cf § 2.2.7) de s'attacher au comment et non au pourquoi.*

(**) *On a Generalised Theory of alternative Inheritence, with Special Reference to Mendel's Law, Proc. Roy. Soc. T 203, Series A (1904) pp. 53-86.*

(***) *Bacon a écrit : Natura rerum magis se prodit per vexationes artis quam in libertate propria : la nature des choses se livre mieux sous des contraintes artificielles que dans sa spontanéité propre.*

De plus à la génétique mendélienne elle-même, l'analyse des données peut s'associer. D'une part les caractères génétiques sont recensés sur des échantillons de population de toute provenance géographique; d'où matière à des analyses révélant les parentés entre ces populations. D'autre part dans les recherches les plus complexes (nous pensons aux groupes tissulaires humains objets des études de M. Greenacre) le généticien doit collationner des informations multidimensionnelles (réactions d'agglutination entre sérums -anticorps- et leucocytes) pour distinguer les locus et établir l'inventaire des allèles : or sa méthode inductive, certes dépendante des principes théoriques venus de Mendel, doit gagner à être complètement formalisée comme un algorithme d'analyse des données.

Ajoutons enfin que c'est à la génétique que la statistique doit un terme qui nous est précieux : celui de facteur. Les études biométriques issues de Galton, bientôt étendues à la psychométrie (mesure des qualités psychiques : intelligence, imagination...) visaient à mettre en évidence un héritage de formes et de manières d'être, régi par ce que d'un nom évocateur on appelait des facteurs. Quand en psychométrie l'analyse des données multidimensionnelles parvint à calculer des grandeurs nouvelles (fonctions des mesures) expressions numériques présumées des tendances profondes de l'individu, ces grandeurs furent nommées facteurs comme si elles étaient la mesure rigoureuse de ce dont on parlait déjà sous ce nom (cf § 2.4) (mais le terme de facteur a été aussi utilisé en génétique pour désigner le caractère élémentaire associé à un locus chromosomique; et en statistique, pour une variable explicitement prise en compte par un plan d'expérience § 2.3.4)

2.2.5. *K. Pearson et le modèle normal* : En 1896 Karl Pearson stimulé par le flot des données biométriques possède enfin l'expression de la densité de la loi normale multidimensionnelle la plus générale en fonction des variances des composantes et de leurs corrélations deux à deux (Edgeworth, dès 1892 avait abordé cette formule). Il généralise la formule de régression trouvée par Galton (cf § 2.2.2) : i.é. donne l'expression de la valeur moyenne $\bar{x}_1$ de x_1, les autres variables $x_2,\ldots, x_n$ étant fixées. De là on passera à la corrélation partielle entre deux composantes : $x_3,\ldots, x_n$ étant fixées, calculer le coefficient de corrélation entre x_1 et x_2 (nous y reviendrons). Par un changement d'axes déjà fait par Bravais pour deux et même trois variables (cf § 2.2.3)(et aussi par Schools; cf Seal in Biometrika T. 54 et Studies) Pearson se ramène à de nouvelles variables indépendantes entre elles; ce qui est poser les calculs d'une analyse factorielle en composantes principales : le titre seul du mémoire : "On lines and planes of closest fit to systems of points in space" (Phil. Mag., T. 2, 1901, pp 559 sqq) suffit à assurer à Pearson la reconnaissance de tous ceux qui aujourd'hui analysent des nuages de points ! (selon C. Burt, in Colloque CNRS 1955, p. 81, on doit ici encore associer à Pearson, Edgeworth); cependant il ne semble pas que le problème du choix d'une métrique dans l'espace ambiant ait été posé, (la norme somme des carrés des coordonnées est acceptée sans discussion). Le premier volume de Biometrika (1901-2) publie sous la signature de W.R. MacDonnell un article d'anthropométrie criminelle (traitant de données recueillies suivant les normes de Bertillon) qui contient selon Burt la première matrice complète de corrélation jamais imprimée. Et MacDonnell, note que Pearson lui a suggéré que de nouvelles coordonnées comptées sur les axes principaux de l'ellipsoïde (ce que nous appelons aujourd'hui des facteurs; mais ce nom ne leur viendra qu'ensuite, de la génétique, cf § 2.2.4) seraient les meilleurs indices pour la reconnaissance des criminels, qu'on puisse déterminer d'après les mesures effectuées. Or le volume 3 de Biometrika offre une étude de K. Pearson "On the inheri-

tence of the mental and moral characters in man, and its comparison with the inheritence of physical characters" où l'appareil des corrélations est appliqué à des mesures de qualités psychiques. De plus dans ses études de la Law of Ancestral Heredity (cf § 2.2.4) Pearson doit décomposer simultanément plusieurs variables normales pour avoir un modèle des corrélations héritées (cf infra § 2.4.2). On voit donc que tous les éléments étaient réunis pour fonder l'analyse factorielle (cf § 2.4) et même l'analyse des correspondances (cf § 2.2.7).

Mais la richesse suggestive du modèle normal multidimensionnel ne doit pas faire oublier que ce modèle est loin d'offrir ne fût-ce qu'en exemples toute la richesse des structures naturelles. Désirant justifier dans sa note historique de 1920 (cf op. laud. in Biometrika T. 13 et Studies) l'introduction de la notion de corrélation partielle peu accessible au non-mathématicien, K. Pearson écrit : ..."trouver la corrélation entre la santé (x_1) d'un enfant et le nombre (x_2) de personnes par pièce en neutralisant les effets de l'âge (x_3) de l'enfant, de la santé (x_4) de ses parents, du salaire (x_5) du père et des habitudes (x_6) de la mère (*), n'est pas un problème moins crucial que celui - reçu de Galton - de la corrélation entre les valeurs d'un même caractère chez parent et enfant". Nous ne contesterons pas l'importance cruciale du problème des corrélations partielles : mais la solution qu'en fournit le modèle normal multidimensionnel est suspecte. Dans ce modèle, (de même que la variance de y pour x fixé est apparue constante à Galton, cf § 2.2.2), la corrélation entre x_1 et x_2 pour $x_3,\ldots,x_n$ fixés ne dépend pas des valeurs assignées à ces dernières variables, parce que, en bref, les sections de la loi normale multidimensionnelle par des plans d'équation $\{x_3 = cte,\ldots, x_n = cte\}$ sont quant au profil toutes égales entre elles à une translation près. Or la pratique de l'analyse des données nous a convaincu qu'autre est la forme véritable des nuages de densité, notamment de ceux qu'offre la sociologie à laquelle K. Pearson emprunte son exemple. Un nuage peut présenter une torsion telle que d'une section à l'autre les corrélations s'inversent; on le voit sur le modèle de nuage tétraédrique, dont les sections se déforment de l'arête AB à l'arête CD (cf fig. 2-3). Et l'on imagine bien que selon les conditions ($x_3,\ldots, x_n$), une même variable x_2 puisse agir positivement ou négativement sur x_1.

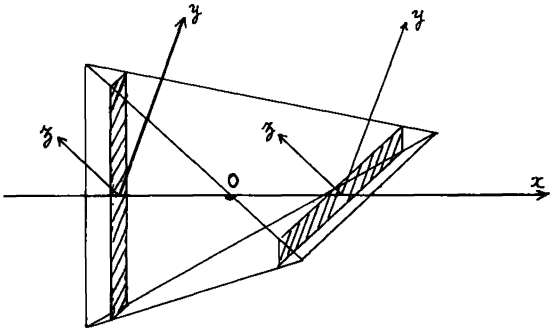

Figure 2-3 : Inversion des corrélations partielles sur les sections d'un nuage tétraédrique par des plans parallèles à deux arêtes opposées à gauche ($x<0$), corrélation positive entre y et z; à droite ($x>0$), corrélation négative entre y et z.

(*) *C'est nous qui ajoutons les coordonnées (x_i), conformément à nos notations.*

Nous nous sommes arrêtés à critiquer le modèle normal. Désirant se faire comprendre des profanes, K. Pearson a choisi un exemple suggestif mais inadéquat; il n'est pas pour autant un fervent inconditionnel de la loi normale (au contraire, cf infra § 2.2.6), loi dont il a bien légitimement considéré les mirifiques symétries. K. Pearson a exploré bien d'autres lois; à la différence de maints statisticiens ses cadets il reconnaissait le primat des données sur les modèles : quoique l'analyse mathématique puisse sous des hypothèses favorables, tirer le meilleur parti d'informations limitées, K. Pearson préférait finalement laisser les élégantes spéculations pour attendre la réponse de données complémentaires. Avouons qu'en analyse des correspondances nous aussi croyons sage de demander à l'expérimentateur et à l'ordinateur ce qu'un statisticien plus laborieux s'efforcerait de tirer de son propre fond mathématique !

2.2.6. L'épreuve du χ^2 : l'observation attentive des lois empiriques conduisit K. Pearson à la découverte de la fameuse épreuve du χ^2, que nous rappellerons ici dans les notations qui nous sont familières. Soit C un ensemble fini (dans la pratique, C pourra être un ensemble discontinu de modalités c d'une qualité; mais initialement, il s'agit d'une suite d'intervalles c en lesquels est découpé assez arbitrairement le champ de variation d'une variable réelle x); $P_C = \{p_c | c \in C\}$ une loi de probabilité sur C; (les p_c sont les probabilités des éventualités c, leur somme est 1); $f_C = \{f_c | c \in C\}$ la loi de fréquence d'un échantillon d'effectif k obtenu par k tirages indépendants à partir de la loi p_C; ($f_c = k(c)/k$, est le rapport au total, du nombre des sorties de l'éventualité c); on calcule l'écart :

$$k||p_C - f_C||^2_{p_C} = \Sigma\{(p_c - f_c)^2/p_c | c \in C\} = k(-1 + \Sigma\{(f_c)^2/p_c | c \in C\});$$

pour k grand, cet écart est distribué comme le carré de la distance à l'origine d'un point de R^n (n = Card C - 1, nombre d'éléments de C moins 1) dont les coordonnées sont normales centrées indépendantes de variance 1. D'où le moyen d'apprécier si une loi empirique f_C déterminée sur un échantillon d'effectif k, diffère significativement d'une loi modèle p_C. Et Pearson se divertit en reprenant les données d'une Pratical verification of the theory of the frequency of errors publiée jadis par Sir George Airy : les données "dont on peut parier à 70 contre 1 (*) qu'elles ne sont pas issues d'une loi normale"; pourtant Airy s'émerveille de leur accord avec la théorie des erreurs et conclut triomphalement "the validity of every investigation in this Treatise is thereby established". De quoi K. Pearson tire cette morale : "Such a passage demonstrates how healthy is the spirit of scepticism in all inquiries concerning the accordance of theory and nature". (nous citons K. Pearson, d'après E.S. Pearson, op. laud. in Biometrika T. 28, 1936).

2.2.7. Causalité et contingence : A cette épreuve de validité ne se borne pas l'usage que fait Pearson de la distance du χ^2. Soient I et J deux ensembles finis; C l'ensemble produit I x J (ou ensemble des

(*) *Pearson veut dire que l'écart (ou χ^2) calculé entre la loi empirique de l'échantillon et le modèle normal n'a qu'une change sur 70 d'être atteint ou dépassé si la loi normale est réellement valide ici. Le sens précis du mot probabilité dans des assertions relatives à la cause d'une observation et en particulier à la forme d'une loi, dépend de la place faite au théorème de Bayes (cf § 1.4.2) et est toujours discuté (cf § 2.3.3).*

couples c associant une modalité ı de I à une modalité ȷ de J); on comparera une loi p_{IJ} au produit $p_I \times p_J$ de ses marges en calculant :

$$d^2 = \| p_{IJ} - p_I \times p_J \|^2 = \Sigma\{ (p_{ij} - (p_i p_j))^2 / (p_i p_j) \mid i \in I, j \in J\}$$

$$= -1 + \Sigma\{ (p_{ij})^2 / (p_i p_j) \mid i \in I, j \in J\}$$

(p_{ij} est la probabilité du couple c = (i,j); $p_i = \Sigma\{p_{ij} \mid j \in J\}$ est la probabilité de i seul etc...). Evidemment cet écart est nul sous l'hypothèse d'indépendance entre i et j : $p_i p_j = p_{ij}$; c'est donc une mesure de la liaison entre deux qualités discontinues ayant I et J pour ensembles de modalité.

Du fait des fluctuations d'échantillonnage, la distance d^2 rigoureusement nulle pour la loi de probabilité sous l'hypothèse d'indépendance, s'écarte de zéro si l'on substitue aux p_{ij}, $p_ı$, $p_ȷ$ des fréquences calculées sur un échantillon d'effectif n; de façon précise la quantité nd^2 est alors distribuée comme un χ^2 à (Card I - 1)(Card J - 1) dimensions : K. Pearson et ses contemporains ne connaissent pas la valeur exacte de ce nombre de dimensions : il était réservé à Fısher de le fixer. Grâce à des conceptions géométrıques qui aboutırent aux théories de l'estimation (§ 2.3.3) et de l'analyse de la variance (§ 2.3.4), Fisher voit dès 1922 (cf. Journal of the Roy. Stat. Soc. T. 85 Pt. 1 pp. 87, 94; 1922 et Contributions) que (pour user d'un langage quı nous est familier) dans la variété de dimension (Card I. Card J)-1 des loıs sur I x J, la sous-variété des lois satisfaisant à l'indépendance ($p_{IJ} = p_I \times p_J$) a pour dimensıon (Card I + Card J - 2) et donc pour codimensions (Card I - 1) x (Card J - 1) (*) : cette codımension est le nombre de degrés de liberté de l'écart entre la loi de l'échantillon et l'hypothèse d'indépendance. Notons toutefoıs que ↳'ignorance du nombre exact des dimensions du χ^2 n'est qu'un détail mineur, relativement aux puissantes conceptions de Pearson sur la contingence et la corrélation.

Ainsi la notion de corrélation, d'abord traduite en nombre par Galton pour les grandeurs numériques est étendue aux variables de toute nature... Reste à relier les deux mesures : reprenons le tableau de correspondance du § 2.2.1 : pour une corrélation r entre x et y,(corrélation de Galton usuelle; ici entre variables normales) on a : $d^2 = \| p_{IJ} - p_I p_J \|^2 = r^2/(1-r^2)$ (telle est du moins la valeur limite, quand s'affinent les subdivisions I et J de l'axe des x et de l'axe des y). C'est ce que calcule Pearson dans son mémoire de 1904 (**); et réciproquement, il définit pour toute correspondance p_{IJ} un coefficient de corrélation généralisé : $r = (d^2/(1 + d^2))^{1/2}$. En analyse des correspondances, on sait que d^2 est la trace, ou somme des valeurs propres (cf 3.5.2).

(*) *Rappelons qu'on appelle codimension d'une variété V considérée dans un espace ambiant E, la différence entre la dimension de E et celle de V.*

(**) *Mathematical contribution to the theory of evolution (XIII) : on the theory of contingency and its relation to association and normal correlation; Drapers'Company Research Memoirs; c'est le mémoire que cite Maung dans un article analysé plus loin (cf § 3.5.2).*

Pearson voit clairement que la relation la plus générale que révèle l'observation des phénomènes naturels est la contingence, la cooccurence; il entrevoit le parti à tirer de l'analyse du tableau rectangulaire recensant ces rencontres; de ce que nous appelons tableau de correspondance (cf § 3.2.4); sans toutefois proposer explicitement la recherche du nombre de dimensions, de facteurs sous-jacents à ces rencontres (*). Il va au delà : comme entre I et J la relation est absolument symétrique, Pearson propose de rejeter la notion polarisée et rigide de cause (i cause j; ou au contraire j cause i) pour se borner aux liaisons réciproques et floues (i et j sont communément unis) qui sont ce qui tombe sous les sens. Lisons "The Grammar of Science" (nous citons la traduction française, faite par Lucien March sur la 3° édition anglaise de 1911) :"La causation est seulement la limite conceptuelle de la corrélation quand la bande (bande centrale du tableau de contingence p_{IJ} : c'est déjà Guttman; cf § 3.4.3) devient si mince qu'elle devient semblable à une courbe" et plus haut "cette courbe est la loi de "causalité" que l'homme introduit dans la nature comme si elle avait une existence réelle. Que représente-t-elle donc ? Une économie de pensée, une routine de perceptions moyenne et approximative..." Il est vrai que l'examen du tableau de Galton ne permet pas de calculer exactement la taille du fils d'après celle du père; ni même de dire laquelle des deux est en quelque sens cause de l'autre ! Mais une réflexion prudemment conduite peut éclairer ce dernier doute ! et l'objectivité de la notion de cause (particulièrement entre termes de niveaux hiérarchiques différents) survit finalement aux plus légitimes critiques. Cependant le statisticien répugnera toujours à accepter selon la boutade de J. Bertrand (cf § 1.6.3) "du cuivre pour de l'or"; c'est à dire un jeu verbal, "nomina quae carent rebus" des noms sans choses, pour l'ordre du réel.

On le voit, pour l'analyse des données de 1975, K. Pearson est un précurseur. Intrépide dans la collecte des données; fécond mais souvent imparfaitement exact dans les constructions mathématiques; plus aventuré encore dans la philosophie; intraitable dans bien des querelles d'Ecole, Karl Pearson a vu son rôle minimisé par la génération des statisticiens élevés dans la doctrine de R.A. Fisher; lequel corrigea mainte inexactitude 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 patronnage de K. Pearson.

(*) *S'il avait eu l'idée de croiser e.g. le couple taille-périmètre thoracique du père avec le même couple de mesures pris sur le fils, (de poser J = ensemble des modalités du couple taille-périmètre chez le père, et de même I chez le fils), K. Pearson n'aurait pas manqué de voir le problème du nombre des dimensions, des facteurs d'une correspondance. D'autres après lui (cf § 3.5.2) écrivirent même explicitement les équations des facteurs sans pénétrer pour autant dons la conception multidimensionnelle.*

Appendice : Note sur la priorité de la découverte de l'inégalité de Fréchet-Darmois-Cramer-Rao; d'après L. Lebart (1).

Un traité récent de statistique (M.G. Kendall et Stuart; T. II. 1962) donne l'énoncé suivant :

soit t un estimateur sans biais d'une fonction $\tau(\theta)$ du paramètre θ d'une Loi; soit $L(x_1, x_2, \ldots, x_n|\theta)$ la densité de probabilité de l'échantillon d'effectif n sur lequel est fondée l'estimation, alors on a l'inégalité suivante :

$$\mathrm{var}(t) = E\{(t-\tau(\theta))^2\} \geq \tau'(\theta)^2/E\{(\partial L/\partial\theta)^2\}$$

et K et S d'ajouter : cette inégalité est dite de Cramer et Rao d'après Cramer (1946) et Rao (1945) mais la priorité semble appartenir à Aitken et Silverstone (1942); avec les références suivantes :

A.C. Aitken et H. Silverstone : on the estimation of statistical parameters, in Proc. Roy. Soc. Edin (A), 62, 369 (1942).

H. Cramer : Mathematical Methods of Statistics; Princeton U.P. (1946)

C.R. Rao Information and accuracy attainable in the estimation of statistical parameters. Bull. Calcutta. Math. Soc. 37, 81-91 (1945).

Aux noms bien connus de Cramer et Rao, et aux deux références communément citées (e.g. par Rao lui-même : cf Linear Statistical inférence and its applications), nous pouvons comme K et S adjoindre des références antérieures à d'autres auteurs, français cette fois, sans oublier le patronage de Sir R. Fisher (*), qu'on omet souvent parce qu'on se lasse de devoir le citer toujours.

Dans la revue de l'Institut international de statistique, M. Fréchet publie en 1943 (pp.185-205), donc avant Cramer et Rao; mais après Aitken, quoiqu'indépendamment de lui, dans l'isolement dû à la guerre, une forme élémentaire où $\tau(\theta) = \theta$; (et où les observations x_i sont indépendantes) de l'inégalité rappelée ci-dessus. Dans la même revue en 1945, G. Darmois, publie un très élégant travail, intitulé "Sur les limites de la dispersion de certaines estimations"), généralisant celui de M. Fréchet au cas où les observations ne sont pas indépendantes, et où le paramètre θ est multidimensionnel.

Cependant le problème auquel G. Darmois donnait ainsi une solution claire et définitive était posé dès 1925 par un travail (Proc. Camb. Phil. Soc. 22 p. 700-725) de R.A. Fisher. Dans un travail de 1935 (J. of the Roy. Stat. Soc. 48, Part I, p. 39), Sir Ronald lui-même affirme que :

1°) l'estimation du maximum de vraisemblance est gaussienne à la limite;

2°) de toutes les estimations gaussiennes, elle est celle d'écart-type minimum.

---

(1) *Chargé de recherches au C.N.R.S. - CREDOC Paris.*

(*) *A l'oeuvre de Sir R.N. Fisher est consacré le prochain article de cette série historique.*

Le deuxième point (qui est un cas particulier de l'inégalité dite de C.R.) a été démontré par D. Dugué en 1936 (cf Comptes Rendus de l'Académie des Sciences; T. 202, pp. 193-195), puis généralisé aux lois limites non normales et au cas de paramètres multidimensionnels (ibid; pp. 452-454). Dans cette note il n'est pas explicitement question de variance d'estimation, mais le résultat n'est pas loin.

Concluons que l'inégalité sous sa forme générale pourrait être attribuée à Darmois-Rao; tandis que Dugué, Aitken, Silverstone et Fréchet marquent entre Fisher et le terme des jalons dignes d'être signalés. H. Cramer, quant à lui, est assurément l'auteur du premier traité démontrant l'inégalité. Il faut pour achever d'éclairer le lecteur (ou peut-être de l'accabler) ajouter que dans le traité de D. Dugué (1958) les noms de Cramer et Rao sont associés à une autre inégalité que celle qui leur est généralement attribuée dans les ouvrages anglo-saxons.

Et l'on conviendra que comme nous l'annoncions au § 2.1, il faut à l'histoire de la statistique "une grande patience dans l'érudition; et... une sainte perspicacité digne du Roi Salomon"; toutes qualités dont nous remercions L. Lebart de nous avoir offert les fruits.