JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

D. WOLKOWITSCH

La géométrie et les diagrammes de la statistique

Journal de la société statistique de Paris, tome 90 (1949), p. 67-74

<http://www.numdam.org/item?id=JSFS_1949__90__67_0>

© Société de statistique de Paris, 1949, tous droits réservés.

L'accès aux archives de la revue « Journal de la société statistique de Paris » (http://publications-sfds.math.cnrs.fr/index.php/J-SFdS) 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/

La géométrie et les diagrammes de la statistique.

Les diagrammes de statistique ont pour objet de mettre en évidence une correspondance entre deux paramètres variables en vue d'établir l'existence d'une relation qui permette de déterminer, quand l'un des paramètres est donné, les limites entre lesquelles peut varier l'autre paramètre.

Lorsque une telle relation existe on dit qu'il y a corrélation. Dans les cas

les plus simples, très nombreux d'ailleurs, les limites sont très rapprochées, elles se confondent même à la limite, on dit alors qu'il existe une liaison fonctionnelle, l'un des deux paramètres est fonction de l'autre.

Tous ces diagrammes pourraient se construire d'une manière unique : dans un système d'axes rectangulaires xOy, on porte l'un des paramètres en abscisses, en considérant des valeurs à intervalles réguliers, aussi rapprochés que possible (nous dirons dans la suite, pour abréger, que l'axe Ox est horizontal). Puis sur chaque verticale A d'abscisse x_A on porte les valeurs observées du paramètre y. Les points A ainsi obtenus correspondent à la valeur x_A du paramètre x.

En opérant de même pour toutes les verticales, nous aurons une série de points figuratifs P_i, de coordonnées $y_i\, x_i$, représentant l'expérience d'indice i, disposés suivant une bande (on dit aussi un nuage). La largeur de cette bande est d'autant plus réduite que la corrélation est plus parfaite; dans le cas de la liaison fonctionnelle la bande devient pratiquement une courbe.

Dans les sciences où les applications des mathématiques sont très développées (astronomie, physique) ainsi que pour de nombreuses questions techniques, on sait souvent, grâce à des considérations théoriques, qu'il existe une liaison fonctionnelle entre les deux paramètres étudiés. Cette liaison se traduit par une courbe qu'il s'agit de déterminer.

En biologie, ainsi que dans toutes les sciences humaines, on ignore, au préalable s'il existe une correspondance entre les deux grandeurs observées. Les diagrammes sont du type décrit plus haut, le problème qui se pose alors est d'établir l'existence d'une corrélation, d'en apprécier le degré de perfection et, le plus souvent, de donner une valeur probable de y qui correspond à une valeur connue de x.

1° *Introduction des notions de l'inertie.*

Les expressions algébriques utilisées en statistique ont un aspect très compliqué; cette complication dissimule des propriétés géométriques élémentaires, simples, et pourtant peu connues et cela qu'il s'agisse de corrélation ou d'ajustement. L'objet de ce travail est de dégager ces propriétés, en nous occupant tout d'abord de l'ajustement. Les expressions algébriques acquièrent de la sorte, une forme concrète qui contribue à leur donner de la clarté.

Nous aurons recours pour cela aux notions usuelles de l'inertie des systèmes plans rigides; il n'y a rien là que de très naturel, si l'on songe qu'à tout moment en statistique, apparaissent les expressions :

$\frac{\Sigma x_i}{N} = X$, $\frac{\Sigma y_i}{N} = Y$ coordonnées du centre de gravité de N masses égales appliquées aux points figuratifs P_i de N expériences, jugées de précision identique; en statistique ce sont des moyennes.

$\Sigma x_i^2 = I'y$, $\Sigma y_i^2 = I'x$ sont les moments d'inertie relatifs aux axes de coordonnées du même système de masses. $\sqrt{\frac{\Sigma x_i^2}{N}}\ \sqrt{\frac{\Sigma y_i^2}{N}}$ sont les rayons de giration correspondants; on les retrouve dans l'écart-type.

$\Sigma x_i y_i = I'xy$ moment centrifuge (ou produit d'inertie) relatif aux axes de

coordonnées, intervient dans les équations des droites de régression. $\sqrt{\frac{\Sigma x_i\, y_i}{N}}$ est le coefficient de Pearson.

Appelons I_x I_y I_{xy} les moments d'inertie et le moment centrifuge relatifs à l'horizontale et à la verticale du centre de gravité un théorème classique concernant les moments d'inertie relatifs à des parallèles permet d'écrire

$$I_x = I'_x - NX^2 \qquad I_y = I'_y - NY^2 \qquad I_{xy} = I'_{xy} - N\,X\,Y.$$

Ainsi voyons nous apparaître la possibilité d'intervention d'une indicatrice d'inertie donnant une représentation de la variation du moment d'inertie relatif à une droite tournant autour d'un point fixe, indicatrice qui devient l'ellipse centrale d'inertie quand le point est le centre de gravité.

2° *Propriétés de l'ellipse centrale d'inertie.*

L'ellipse que nous emploierons ici sera exclusivement l'ellipse de Culmann; nous la désignerons par la lettre *e*. Elle est concentrique et homothétique à l'ellipse centrale classique qui est celle de Poinsot-Lamé, elle possède de nombreuses propriétés intéressantes; nous nous bornerons à rappeler les trois suivantes qui nous semblent les plus intéressantes :

I. Étant donnés dans un plan, des points P_i affectés de masses m_i, et d_i leurs distances à une droite D de ce plan δ, les quantités $m_i\, d_i = \mu_i$ considérées comme de nouvelles masses, appliquées aux points P_i, admettent un centre de gravité qui est dans l'ellipse *e*, l'antipôle de la droite D.

Pôle et antipôle sont deux points symétriques l'un de l'autre par rapport au centre, sur le diamètre conjugué de la droite D.

II. On déduit de cette propriété fondamentale que l'ellipse *e* est l'enveloppe des droites dont les distances au centre de gravité G sont égales aux rayons de giration relatifs aux diamètres parallèles. C'est cette propriété qui fait de l'ellipse *e* une indicatrice d'inertie, car elle donne une représentation de la variation du moment d'inertie relatif aux droites passant par le centre de gravité.

III. Considérons un diamètre HH′ variable de l'ellipse *e*. Désignons par h_i la distance du point P_i à ce diamètre, distance mesurée parallèlement à une direction fixe O*y* par exemple. La somme $\Sigma m_i h_i^2$ passe par un minimum lorsque HH′ est le diamètre conjugué de O*y* dans l'ellipse *e* (nous dirons pour abréger que HH′ est le d.c.v. (diamètre conjugué de la verticale).

La propriété I dispense de déterminer au préalable les éléments de l'ellipse *e*, quand on veut construire ce d. c. v. Celui-ci passe en effet par le centre de gravité G, des masses *m*, en outre, il passe par l'antipôle γ de la verticale O*y*, centre de gravité des masses μ_i appliquées aux points P_i. La droite est la droite cherchée.

On déduit également de cette propriété I que le moment centrifuge est nul quand les deux droites sont anticonjuguées dans l'ellipse *e*, (l'une contient l'antipôle de l'autre). En particulier le moment centrifuge relatif à deux diamètres conjugués de *e* est nul.

3° *Applications.*

Prenons un diagramme du type biologique. Soit A_i les points figuratifs p portés par la verticale A, d'abscisse x_A, ils représentent les expériences correspondant à la valeur x_i du paramètre x. Leurs ordonnées sont désignées par y_i leur nombre par n_A.

L'ordonnée du centre de gravité A_0 des points A_i est

$$A_0 = \frac{\Sigma y_i}{n_A}.$$

Ce point jouit d'une propriété classique : A' représentant un point fixe de la verticale, la somme $\Sigma \overline{A' A_i}^2 = \Sigma (y' - y_i)^2$ passe par un minimum quand le point A' coïncide avec le point A_0. Appelons $A_0, B_0, C_0 \ldots$ les centres de gravité des verticales, ils forment un polygone tel que la somme des carrés des distances verticales h_i^2 soit minima.

Quand les points $A_0 B_0 \ldots$ sont alignés, la droite qui les contient est une droite de régression.

La propriété III du paragraphe 3° donne une construction immédiate de la droite sans qu'il soit besoin de passer par l'ellipse e ni par les centres de gravité $A_0 B_0 \ldots$ pour la déterminer. C'est le d.c.v. dans l'ellipse centrale d'inertie des points P_i. Elle passe par le centre de gravité G : $X = \frac{\Sigma x_i}{N}$ $Y = \frac{\Sigma y_i}{N}$ avec $N = n_A + n_B + \ldots$

et aussi par l'antipôle γ de Oy dans l'ellipse e, γ est le centre de gravité des masses $\mu_i = x_i$; ses coordonnées sont donc

$$\xi = \frac{\Sigma x_i x_i}{\Sigma x_i} = \frac{\Sigma x_i^2}{\Sigma x_i} \qquad \eta = \frac{\Sigma x_i y_i}{\Sigma x_i}.$$

L'équation de la droite s'écrit sous la forme commode d'un déterminant :

$$\begin{vmatrix} x & y & 1 \\ X & Y & 1 \\ \xi & \eta & 1 \end{vmatrix} = \begin{vmatrix} x & y & 1 \\ \Sigma x_i & \Sigma y_i & N \\ \Sigma x_i^2 & \Sigma y_i x_i & \Sigma x_i \end{vmatrix} = 0$$

ou si l'on préfère

$$\frac{x - X}{\xi - X} = \frac{y - Y}{\eta - Y}$$

ou encore

$$y = x \frac{\eta - Y}{\xi - X} + \frac{Y\xi - X\eta}{\xi - X}.$$

Le paramètre directeur est

$$\alpha = \frac{\Sigma y_i \, \Sigma x_i - N \Sigma x_i y_i}{(\Sigma x_i)^2 - N \Sigma x_i^2}$$

l'ordonnée à l'origine

$$\beta = \frac{\Sigma x_i^2 \, \Sigma y_i - N \Sigma x_i y_i \, \Sigma x_i}{N \Sigma x_i^2 - (\Sigma x_i)^2}.$$

La relation entre l'ordonnée du centre de gravité des points d'une verticale A et l'abscisse de cette verticale sera donc $a = \alpha\, x_A + \beta$.

Mais la direction verticale n'est pas privilégiée; on peut tout aussi bien grouper les points P_i suivant des horizontales. On obtient des centres de gravité A'_0, B'_0.,.. conduisant, s'ils sont alignés, à une deuxième droite de régression, qui sera le diamètre conjugué de l'horizontale dans la même ellipse *e*. Ce diamètre passe par le centre de gravité G' G et par l'antipôle de l'horizontale. Le paramètre directeur peut s'écrire directement à partir de α en intervertissant les lettres x et y, il vient

$$\frac{\Sigma x_i\ \Sigma y_i - N\ \Sigma x_i\, y_i}{(\Sigma y_i)^2 - N\ \Sigma y_i^2}.$$

Les deux droites de régression diffèrent en général, puisqu'elles passent toutes deux par le centre de gravité G et que leurs paramètres directeurs diffèrent. Leur angle caractérise le degré de corrélation.

4° *Cas de la liaison fonctionnelle.*

L'aspect du problème se modifie très légèrement; les points figuratifs P_i sont en nombre assez réduit, ils sont répartis suivant une bande étroite, d'allure rectiligne (nous généraliserons plus loin); il s'agit de trouver la droite qui rende minimum l'expression

$$\Sigma\, d_i^2$$

somme des carrés des écarts verticaux. Les points P_i jouent ici le rôle des centres de gravité des verticales A, B, C... Le problème revient à déterminer les deux paramètres qui définissent la droite; c'est cette détermination qui constitue l'*ajustement de la droite* $y = ax + b$ *au diagramme des points* P_i.

La droite cherchée est le d.c.v. dans l'ellipse centrale d'inertie, *e*, du système des points P_i (propriété III du § 2).

Le diamètre conjugué de l'horizontale dans la même ellipse rendra minimum la somme des carrés des écarts horizontaux.

Si la représentation linéaire de la corrélation est appropriée, les deux diamètres sont assez voisins l'un de l'autre; leur angle indique dans quelle mesure on est fondé à adopter une loi linéaire $y = \alpha x + b$ pour cette représentation.

Les deux diamètres coïncideraient si la liaison fonctionnelle était rigoureusement linéaire.

5° *Généralisation.*

Quand la loi linéaire est insuffisante il est naturel d'essayer une loi de la forme

$$y = a\, f(x) + b$$

$f(x)$ est une fonction dont le choix est dicté par la comparaison de la forme de la courbe $y = f(x)$ avec celle du nuage des points P_i; a et b sont deux paramètres à déterminer de manière que la somme des carrés des écarts soit minima. La détermination de a et b constitue encore l'*ajustement* de la courbe aux résultats des expériences représentées par les points P_i.

Nous ramenons le problème au précédent en opérant le changement de variable $X = f(x)$, $Y = y$ qui fait correspondre à un point P_i (x_i et y_i) un point

$$Q_i \quad X_i = f(x_i) \text{ et } Y_i.$$

Tout système de valeurs a et b donne une courbe $y = af(x) + b$ dans le plan des P_i et une droite correspondante dans le plan des Q_i : $Y = aX + b$. L'écart vertical du point Q_i à la droite est $Y_i - aX_i - b$ celui du point P_i à la courbe : $y_i - af(x_i) - b$ ces deux écarts sont égaux en vertu du changement de variable, les sommes de leurs carrés seront donc égales elles passeront donc simultanément par un minimum; en d'autres termes, la courbe C qui donne le minimum de cette somme dans le plan des P_i correspond à la droite D qui dans le plan des Q_i donne le minimum. Nous appellerons cette droite, droite auxiliaire.

Nous savons construire cette droite D, c'est le d.c.v. dans l'ellipse centrale d'inertie du système des points Q_i. Son équation sous forme de déterminant est

$$\begin{vmatrix} X & Y & 1 \\ \Sigma X_i & \Sigma Y_i & N \\ \Sigma X_i^2 & \Sigma X_i Y_i & \Sigma X_i \end{vmatrix} = 0$$

on en déduit celle de la courbe cherchée dite courbe de regression :

$$\begin{vmatrix} f(x) & y & 1 \\ \Sigma f(x_i) & \Sigma y_i & N \\ \Sigma [f(x_i)]^2 & \Sigma y_i f(x_i) & \Sigma f(x_i) \end{vmatrix} = 0$$

ou

$$y.[(\Sigma f(x_i))^2 - N.\Sigma f(x_i)^2] = f(x)[\Sigma y_i \Sigma f(x_i) - N \Sigma y_i f(x_i)] + \Sigma f(x_i) \Sigma y_i f(x_i) - \Sigma y_i \Sigma f x_i)^2]$$

Exemple :

Soit à ajuster une courbe de la forme :

$$y = \frac{a}{x} + b$$

à un diagramme de quatre expériences dont les points figuratifs ont pour coordonnées les nombres des colonnes 1 et 2 du tableau ci-après :

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
x_i	y_i	$X_i = \frac{1}{x_i}$	X_i^2	$X_i y_i$	$Z_i = x_i y_i$	Z_i^2	$Z_i x_i$
1	12	1	1	12	12	144	12
2	7	0,5	0,25	3,5	14	196	28
5	4	0,2	0,04	0,8	20	400	100
10	3	0,1	0,01	0,3	30	900	300
18	26	1,8	1,3	16,6	76	1640	440

coordonnées du centre de gravité des points $X_i\ y_i$

$$\frac{\Sigma X_i}{4} \qquad \frac{\Sigma y_i}{4}$$

coordonnées de l'antipôle de Oy dans l'ellipse centrale d'inertie du système des points X_i, y_i

$$\frac{\Sigma X_i^2}{\Sigma X_i} \qquad \frac{\Sigma X_i y_i}{\Sigma X_i}$$

équation de la droite auxiliaire

$$\begin{vmatrix} X & y & 1 \\ 1{,}8 & 26 & 4 \\ 1{,}3 & 16{,}6 & 1{,}8 \end{vmatrix} = 0$$

ce qui donne pour la courbe de régression

$$0 = \frac{1}{x} \cdot 19{,}6 - y.\ 1{,}96 + 3{,}92 \text{ ou} \qquad y = \frac{10}{x} + 2.$$

L'équation de la courbe $y = \frac{a}{x} + b$ peut s'écrire, en chassant le dénominateur, $xy = a + bx$ de la forme $x = \alpha xy + \beta$. Nous posons $Z = xy$; ce changement de variables nous conduit à une deuxième droite auxiliaire $x = \alpha Z + \beta$ qui rend minimum la somme des carrés des écarts *horizontaux :*

$$\Sigma (x_i - \alpha Z_i - \beta)^2 \text{ minimum}$$

cette droite est, dans l'ellipse centrale d'inertie, des points de coordonnées x_i et Z_i, le diamètre conjugué de la droite Ox. Le centre de gravité a pour coordonnées $\frac{\Sigma x_i}{4}\ \frac{\Sigma Z_i}{4}$ l'antipôle de Ox

$$-\frac{\Sigma Z_i x_i}{\Sigma Z_i} \qquad \frac{\Sigma Z_i^2}{\Sigma Z_i}.$$

Les Σ intervenant dans l'équation se trouvent aux colonnes (6) (7) et (8) du tableau. L'équation de la droite auxiliaire est

$$0 = \begin{vmatrix} x & Z & 1 \\ \Sigma x_i & \Sigma Z_i & 4 \\ \Sigma x_i Z_i & \Sigma Z_i^2 & \Sigma Z_i \end{vmatrix} = \begin{vmatrix} x & Z & 1 \\ 18 & 76 & 4 \\ 440 & 1640 & 76 \end{vmatrix}$$

conduisant à une courbe de régression qui, après réductions, s'écrit

$$49\ xy = 490 + 98\ x$$

ou

$$y = \frac{10}{x} + 2.$$

C'est *la même hyperbole que précédemment.*

Dans le cas général les deux courbes seraient distinctes. Leur coïncidence, dans le cas envisagé, résulte de ce que les quatre points $x_i\ y_i$ se trouvent sur une hyperbole dont l'équation a la forme indiquée.

Nous vérifions ainsi que s'il existe une courbe de la forme choisie C, $y = a f(x) + b$ qui contienne tous les points P_i sans qu'on le sache, la méthode suivie conduira à l'équation de la courbe C. On ne risque pas de la manquer au cours des calculs. Le raisonnement suivant montre d'ailleurs que, dans ce cas particulier, il ne peut en être autrement.

La somme des carrés des écarts Σ, est une somme de termes tous positifs, son minimum est donc O, et cette valeur est atteinte quand tous les termes sont nuls.

Tous les points P_i appartenant à la courbe C, par hypothèse, la somme des carrés des écarts relative à cette courbe est nulle.

Les opérations ne peuvent conduire à une courbe C′ différente de C, dont l'équation serait $y = a' f(x) + b'$ car la somme Σ_c, serait $> o$ c'est-à-dire $\Sigma'_c > \Sigma_c$ ne serait pas un minimum.

D. Wolkowitsch.