MATHÉMATIQUES ET SCIENCES HUMAINES

M. GIRAULT

Liaisons, corrélation, régression

Mathématiques et sciences humaines, tome 5 (1964), p. 11-22

<http://www.numdam.org/item?id=MSH_1964__5__11_0>

© Centre d'analyse et de mathématiques sociales de l'EHESS, 1964, tous droits réservés.

L'accès aux archives de la revue « Mathématiques et sciences humaines » (http://msh.revues.org/) 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/

M. GIRAULT

LIAISONS, CORRELATION, REGRESSION

INTRODUCTION

1.1.- La physique classique a utilisé la notion de liaison fonctionnelle entre deux ensembles de grandeurs:

Exemple: Soit une barre métallique portée à différentes températures: t_1 t_2 .... t_n: elle prend les longueurs l_1 l_2 .... l_n.

Chaque fois que la barre est portée à la température t_2 par exemple, elle prend la longueur l_2. Ainsi à toute température (d'un certain intervalle de températures) est associée une longueur et une seule. Telle est la notion de fonction ou de liaison fonctionnelle $t \longrightarrow l$.

Ce type de modèle, bien adapté aux phénomènes étudiés en physique classique ne convient généralement pas dans les sciences biologiques, économiques, humaines.

1.2.- L'idée de liaison en probabilité a été introduite pour la première fois par Francis Galton dans ses travaux sur l'hérédité dans les espéces végétales puis animales. Les principaux résultats furent publiés en 1889 sous le titre "Natural inheritance".

L'auteur s'intéresse tout spécialement aux mesures X et Y d'une même partie du corps chez deux êtres dont l'un est fils de l'autre. Appelons "taille" la mesure retenue:

$$\left.\begin{matrix} X_i \\ Y_j \end{matrix}\right\} \text{ est un couple de tailles observées } \left\{\begin{matrix} X_i & \text{celle de l'individu (i)} \\ Y_j & \text{" " (j)} \end{matrix}\right.$$

où (j) est fils de (i).

Galton montre que la loi de fréquence des couples de nombres (X, Y) est bien représentée par une loi de distribution à deux dimensions généralisant celle de Laplace-Gauss et étudiée précédemment par Bravais(1) (Mémoire de 1846).

Le modèle utilisé par Galton est très fréquemment employé. Ce modèle présente des particularités qu'il faut bien comprendre et, pour cela, il importe de distinguer plusieurs notions fondamentales concernant la liaison en probabilité, notions qui dans le modèle particulier de Galton se trouvent soit confondues soit liées entre elles d'une manière particulière. En d'autres termes, il faut bien distinguer les propriétés générales de toute distribution des propriétés particulières de la distribution de Laplace-Gauss à plusieurs dimensions.

---

(1) Bravais ne semble pas s'être intéressé à la liaison en probabilité.

Remarque préliminaire: Les notions que nous allons étudier se rattachent à des distributions sur un espace à plusieurs dimensions (tout spécialement deux): ce sont des distributions sur des couples (X, Y) de nombres.

Ces distributions peuvent être indifféremment des distributions statistiques ou des distributions de probabilité; ces dernières elles-mêmes pouvant être discrètes ou continues.

a) Distribution statistique: échantillon de taille (n): ensemble de n couples (x_1, y_1); $x_2\ y_2)$, ..., (x_n, y_n) chacun affecté de la "masse" $\frac{1}{n}$

b) Distribution de probabilité

b.1 - Discrète: ensemble de couples (x_i, y_i)

$$x_i \in \{x_1, x_2, \ldots, x_n, \ldots\}$$

où

$$y_j \in \{y_1, y_2, \ldots, y_m, \ldots\}$$

le couple (x_i, y_j) étant affecté de la probabilité P_{ij}

b.2 - Continue: ensemble de couples (x,y), où x et y appartiennent à des ensembles continus (intervalles où même $\mathbf{R}$).

Chaque pavé $\begin{Bmatrix} (x,\ x + \Delta x) \\ (y,\ y + \Delta y) \end{Bmatrix}$ est affecté d'une probabilité.

En particulier tout pavé élémentaire $\begin{Bmatrix} (x,\ x + dx) \\ (y,\ y + dy) \end{Bmatrix}$ est affecté de la probabilité $f(x, y)\ dx\ dy$.

Dans chaque cas, nous choisirons pour fixer les idées un type de distribution tantôt discret, tantôt continu; étant bien entendu que les notions étudiées se définissent pour les deux types de distributions.

2.- LIAISON EN PROBABILITE (ou liaison stochastique)

2.1.- Considérons donc une épreuve à laquelle est associé le couple de nombres (x, y) distribués, par exemple, suivant une loi discrète.

$$\begin{cases} x \in \{x_1, x_2, \ldots, x_n\} = A \\ y \in \{y_1, y_2, \ldots, y_m\} = B \end{cases} \quad (x, y) \in A \times B$$

Posons

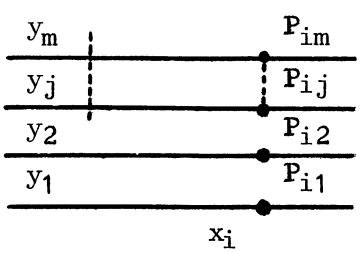
$$P_{ij} = P(X = x_i, Y = y_i) \qquad i = 1, \ldots n;\ j = 1, \ldots m$$

Rappelons les définitions des lois marginales et des droits conditionnelles:

Distribution marginale de X: soit $a_i = P(X = x_i)$ $i = 1, \ldots n$ On a, (axiome d'additivité):

$$a_i = \sum_j P_{ij}$$

que l'on note encore P_i.



La donnée des x_i et celle des $p_{i.}$ définit $(p_{i.})$ la distribution marginale de X.

Distribution marginale de Y : on a de façon analogue

$$b_j = p_{.j} = \sum_i p_{ij}$$

Les y_j et les $p_{.j}$ définissent la distribution marginale de Y.

Remarque: Si la distribution est continue, de densité superficielle f(x, y). Les densités marginales sont :

$$a(x) = \int f(x, y)\, dy \quad \text{pour X}$$

$$\text{et } b(y) = \int f(x, y)\, dx \quad \text{pour Y}$$

ou, si l'on préfère, les lois élémentaires sont respectivement

pour X $\quad a(x)\, dx = \int_y f(x, y)\, dx\, dy$

pour Y $\quad b(y)\, dy = \int_x f(x, y)\, dx\, dy$

2.2.- Distribution conditionnelle de Y si $X = x_i$

On effectue l'épreuve donnant le couple (X, Y) et l'on sait que $X = x_i$. La probabilité d'avoir $Y = y_j$ est alors

$$b_{(j)}^{(i)} = \frac{p_{ij}}{p_{i.}} \quad \text{notée} \quad \text{Prob.} \left\{ Y = y_j \,/\, X = x_i \right\}$$

On aurait de façon analogue:

$$\text{Prob.} \left\{ X = x_i \,/\, Y = y_j \right\} = a_i^{(j)} = \frac{p_{ij}}{p_{.j}}$$

Et, dans le cas d'une distribution continue:

loi élémentaire de Y si $X = x$:

$$b_{(y)}^{(x)}\, dy = \frac{f(x,y)\, dx\, dy}{a(x)\, dx} = \frac{f(x,y)}{a(x)}\, dy$$

2.3.- Dépendance en probabilité

On peut considérer non seulement les lois conditionnelles de Y si X prend une valeur particulière.

$$\left[\begin{array}{l} X = x_i \quad \text{dans le cas discret} \\ x < X < x + dx \quad \text{dans le cas continu} \end{array} \right]$$

mais les lois conditionnelles de Y lorsque X appartient à un sous-ensemble de valeurs possibles: un intervalle (x_0, x_1) par exemple.

On dit que Y dépend de X (en probabilité) si les lois conditionnelles de Y dépendent des conditions imposées à X. Il y a liaison en probabilité si

$$b_{(j)}^{(i)} \text{ dépend de } a_i$$

2.4.- Indépendance en probabilité

Au contraire on dit de Y qu'il est indépendant de X (en probabilité) si toute loi conditionnelle de Y est indépendante des conditions imposées à X; c'est-à-dire si toutes les lois conditionnelles sont identiques.

On aura en particulier:

(distribution discrète): $b_{(j)}^{(1)} = b_{(j)}^{(2)} = \ldots = b_{(j)}^{(n)}$ et cela quel que soit j.

soit :

$$\frac{p_{1J}}{p_{1\cdot}} = \frac{p_{2j}}{p_{2\cdot}} = \frac{p_{3j}}{p_{3\cdot}} = \ldots\ldots = \frac{p_{nj}}{p_{n\cdot}} = \frac{\sum_k p_{kj}}{\sum_k p_{k\cdot}} = p_{\cdot j}$$

d'où :

$$\forall i, \forall j : \qquad \boxed{p_{ij} = p_{i\cdot}\, p_{\cdot j}} \qquad (1)$$

De la même manière, dans le cas continu, on démontre qu'il faut avoir, quels que soient x et y

$$\boxed{f(x,y) = a(x) \cdot b(y)} \qquad (2)$$

Des relations (1) et (2) on déduit plusieurs conséquences. On a simplement supposé, pour obtenir ces relations, que les lois conditionnelles de Y pour des conditions élémentaires de X $\left[X = x \text{ ou } x < X < x + dx\right]$ étaient indépendantes de x :

On déduit de (1) ou de (2)

- Que toute loi conditionnelle de Y pour des conditions exprimées sur X est identique à la loi marginale de Y,
- Que toute loi conditionnelle de X pour des conditions exprimées sur Y est identique à la loi marginale de X.

On exprime cette situation en disant que les v.a. X et Y sont mutuellement indépendantes en probabilité.

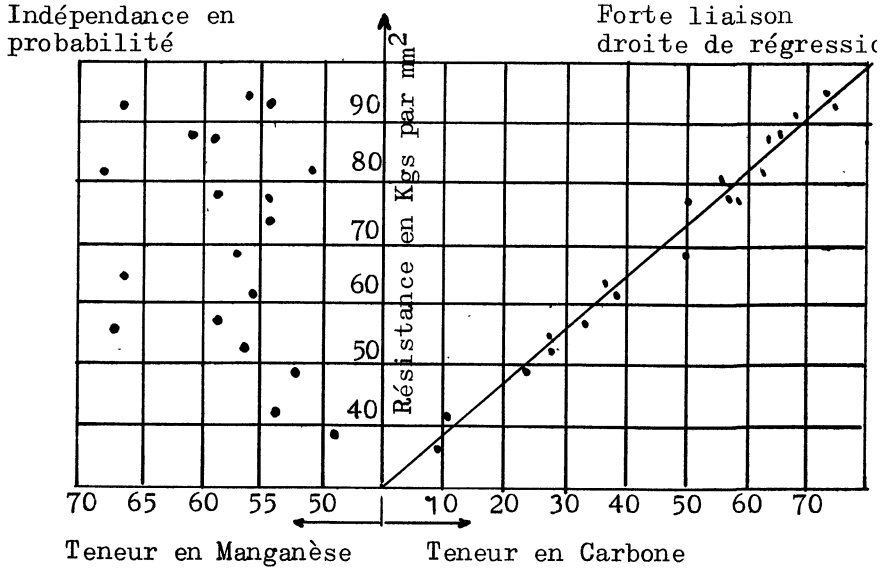
Dans ce cas, la distribution est complètement déterminée par la donnée des distributions marginales.

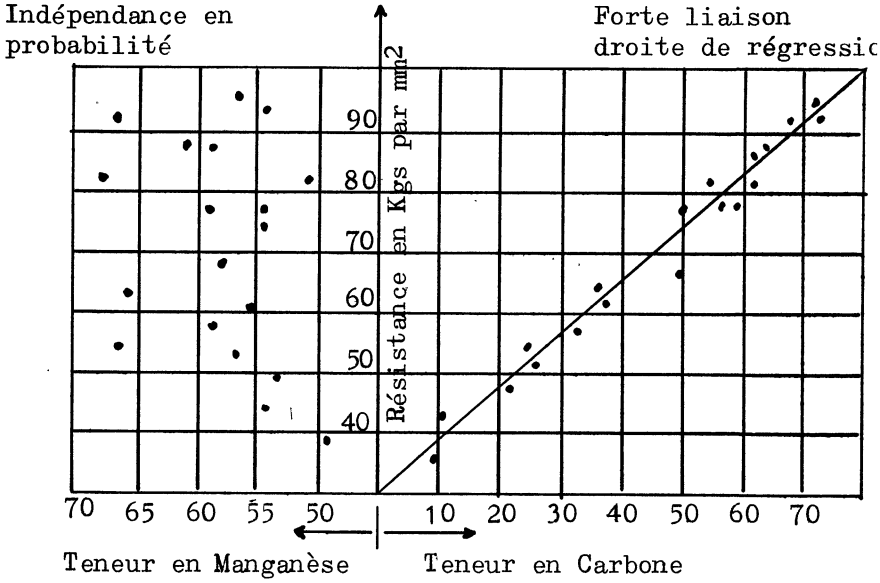
Une connaissance quelconque sur X, ne modifie en rien l'information que l'on possède sur Y.

3.- MOMENTS DU 1er ORDRE - Ligne de régression

Pour varier les notations, nous supposerons ici que la distribution est continue, de densité superficielle f(x,y).

Illustration:





La loi élémentaire de Y

$$\text{si } X = x \qquad \text{est } b^{(x)}_{(y)}\, dy = \frac{f(x,y)}{a(x)}\, dy$$

où $a(x) = \int_y f(x\ y)\, dy$

La moyenne de (Y si X = x) est (quand elle existe)

$$m^{(x)} = y(x) = \int_y y\, b^{(x)}_{(y)}\, dy$$

$\overline{y}\ (x)$ est une fonction de x; sa courbe figurative est dite <u>ligne de régression</u> de Y en x.

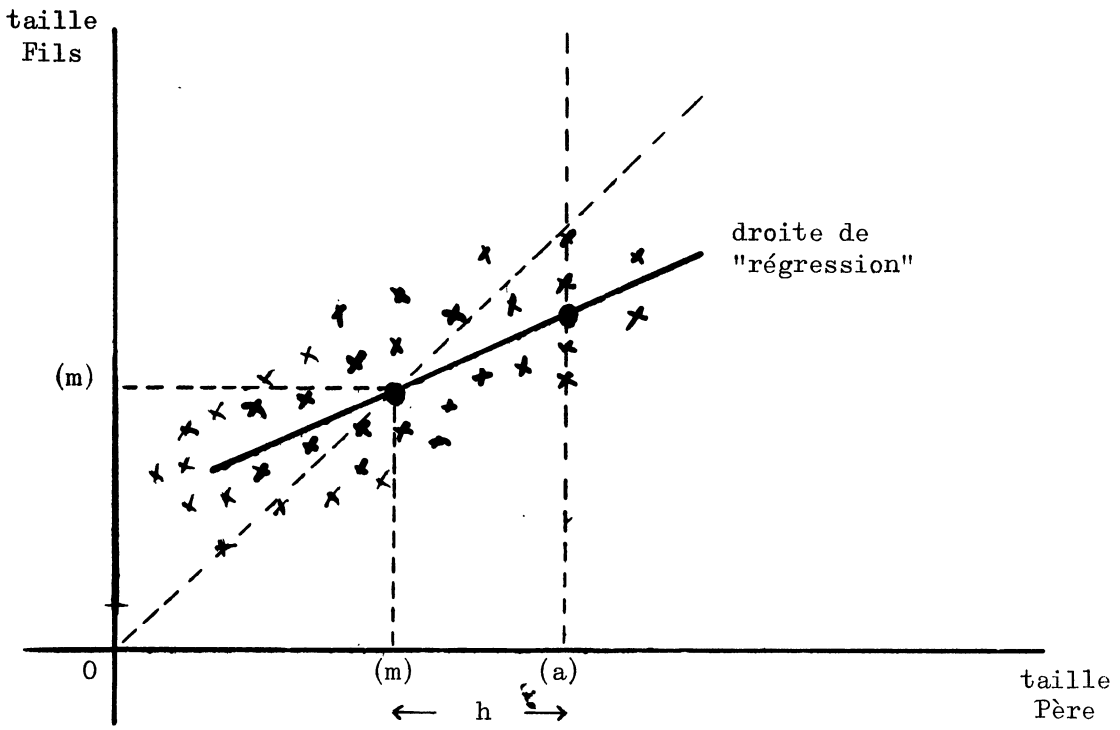
On définit d'une manière analogue la ligne de régression de X en y; c'est le lieu des moyennes conditionnelles de X pour Y = y.

Si les v.a. X et Y sont indépendantes, les lignes de régression sont des droites parallèles aux axes.

Une ligne de régression (de Y en x par exemple) est un indice de la liaison en probabilité de Y avec le "facteur" X. L'origine du terme "régression" est la suivante.

Origine du terme "regréssion"

Etude de l'hérédité des tailles (F. GALTON puis K. PEARSON).



Soit Y la taille d'un individu, X celle de son père. Les couples (X,Y) sont décrits par une loi de probabilité où la ligne de régression de Y en x est une droite de pente r positive inférieure à 1.

Soit m la taille moyenne de la population étudiée. Les "grands" individus (de taille a = m + h par exemple) ont des fils dont les tailles sont en moyenne: a + rh. Ces fils sont des individus (en moyenne) grands mais ce caractère "grand" est moins accusé qu'il ne l'est chez leurs pères: on dit que ce caractère est en régression d'une génération à la suivante.

4.- MOMENTS DU SECOND ORDRE

4.1.- Pour décrire la liaison des grandeurs X et Y, les indices donnés par les moments du premier ordre sont insuffisants. On s'intéresse donc aux moments du second ordre (en supposant là encore qu'ils existent).

Soit $\mu_{20} = \sigma_x^2$ la variance de X

$\mu_{02} = \sigma_y^a$ la variance de Y

$\mu_{11} = E\left[(X-\overline{X})\ (Y-\overline{Y})\right]$

} moments du 2ème ordre du couple X, Y.

Soit enfin $\left[\sigma_{(y)}^{(x)}\right]^2$ la variance conditionnelle de Y / si X = x.

On appelle <u>fonction scédastique</u> (la fonction $z(x) = \sigma_{(y)}^{(x)}$. Si $\sigma_{(y)}^{(x)} = c^{te}$ on dit qu'il y a <u>homoscédasticité</u> de Y par rapport à x.

4.2.- Indices de corrélation

Plusieurs indices ont été proposés, mais en fait le plus utilisé est le <u>coefficient de corrélation</u>

$$\rho\ (x,\ y) = \frac{\mu_{11}}{\sigma_x\ \sigma_y}$$

La signification du coefficient de corrélation s'interprète commodément sur l'ellipse des variances que nous définirons plus loin.

Un autre indice intéressant est le rapport de corrélation que nous allons maintenant définir.

4.3.- Analyse de variance - Rapport de corrélation

Sur le couple aléatoire (X,Y) on considère:

la moyenne générale de Y, soit $E(Y) = \overline{y} = \int\int yf(x\ y)\ dx\ dy$

les moyennes conditionnelles de Y si X = x

$$\text{soit } E\left[{}^{Y}/_{X} = x\right] = \overline{y}_{(x)} = \int yf(x,y)\ dy.$$

Soit (x,y) un couple de valeurs obtenues.

L'écart de y peut se décomposer de la manière suivante:

$$y - \overline{y} = (y - \overline{y}_x) + (\overline{y}_x - \overline{y}) \qquad (1)$$

La variance de Y est $\sigma^2 = E\left[(y - \overline{y})^2\right] = \int\int(y - \overline{y})^2\ f(x,y)\ dx\ dy$

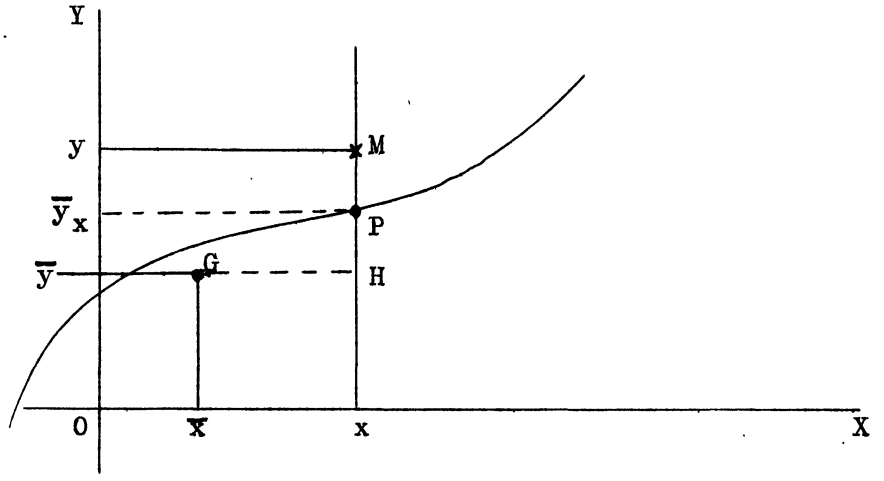
En remplaçant $(y - \overline{y})$ par sa valeur (1) on obtient:

$$\sigma^2 = \int\int \left[(y - \bar{y}_x) + (\bar{y}_x - \bar{y})\right]^2 \; f(x,y)\,dx\,dy$$

$$= \int\int (y - \bar{y})^2 \; f(x,y)\,dx\,dy + 2\iint (y - \bar{y}_x)(\bar{y}_x - \bar{y}) \; f(x,y)\,dx\,dy + \int\int (\bar{y}_x - \bar{y})^2 \; f_{x,y}\,dx\,dy$$

variance totale $= \sigma^2$ = variance conditionnelle (ou résiduelle) σ_2^2 + nulle + variance liée (2) σ_1^2

La variance conditionnelle (σ_2^2) mesure la dispersion autour de la ligne de régression. Elle est nulle si la distribution est entièrement sur la courbe de régression et dans ce cas seulement.



au couple de valeurs (x,y) représenté par M sont associés:
HM = écart total
PM = écart conditionnel
HP = écart lié.

La variance liée (σ_1^2) mesure la dispersion des moyennes conditionnelles $\bar{y}_x$. Elle est nulle si la courbe de régression est une droite horizontale, et dans ce cas seulement.

Rapport de corrélation (PEARSON - 1905).

On appelle rapport de corrélation de Y en X le nombre positif R tel que

$$R^2 = \frac{\sigma_1^2}{\sigma^2} = 1 - \frac{\sigma_2^2}{\sigma^2}$$

En vertu de l'équation (2) on voit que $0 < R < 1$

$R = 0 \iff \bar{y}_{(x)} = c^{te}$ (la courbe de régression est une droite horizontale)

$R = 1 \iff$ distribution sur la courbe de régression donc liaison fonctionnelle entre X et Y.

4.4.- Droite des moindres carrés

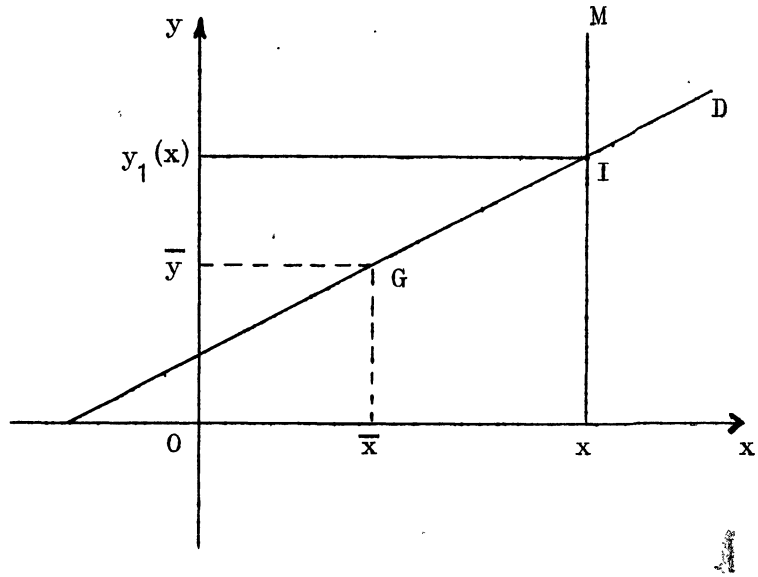
On appelle droite des moindres carrés (de Y/x) la droite D telle que l'espérance mathématique de $\overline{IM}^2$ soit minimum :

Son équation est $y_1(x)$ tel que $\iint\left[y-y_1(x)\right]^2 f(x,y)dxdy$ soit minimum.

La recherche de ce minimum montre que D passe par G, centre de la distribution. D a pour pente

$$\rho \frac{\sigma_y}{\sigma_x}$$

<u>Remarque</u>: Certains auteurs appellent la droite D: "droite de régression" de Y en x.



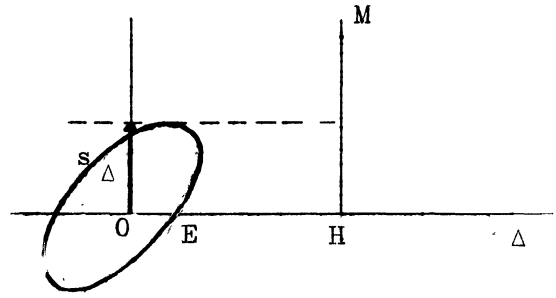
4.5.- Ellipse des variances

Toutes les notions précédentes: 4-1, 4-2, 4-3 et 4-4 se rattachant aux moments du 2ème ordre. Les moments des formes linéaires en X, Y (expressions de la forme aX + bY) s'expriment facilement à partir de l'ellipse des variances.

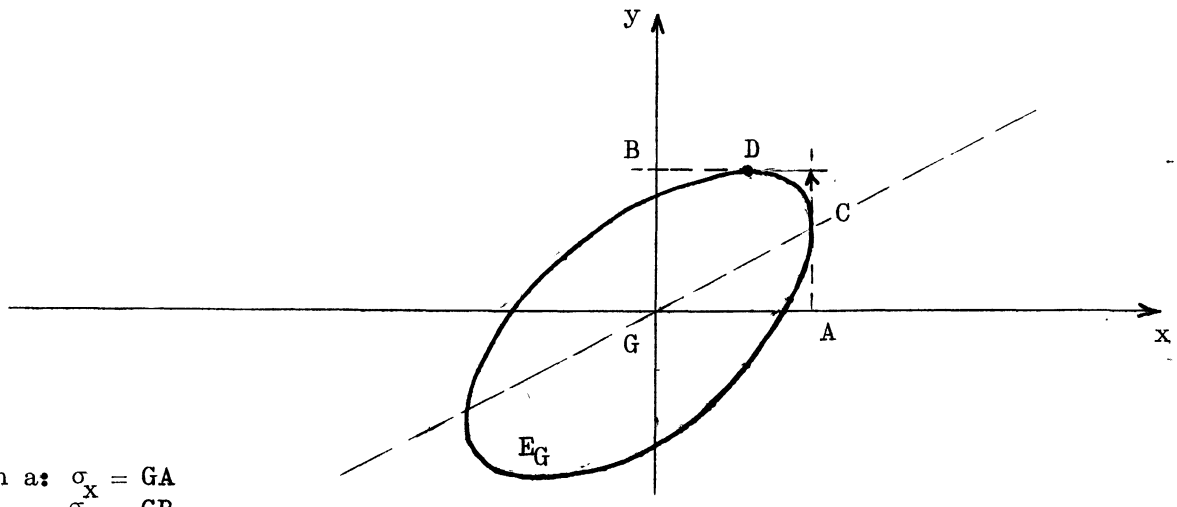
Soit une distribution sur le plan (ox,oy) admettant des moments du second ordre: $E(X^2)$ et $E(Y^2)$ (donc aussi $E(X.Y)$).

Il existe une ellipse E de centre O qui a la propriété suivante:

Soit Δ une droite issue de l'origine. Le moment de la distribution par rapport à Δ (espérance mathématique $\overline{MH^2}$) est égal à s_Δ^2 ou s_Δ est la distance de O à la tangente à E parallèle à Δ.



En particulier, si l'on rapporte la distribution à son centre G l'ellipse associée: E_G est dite "<u>ellipse des variances</u>"



On a: $\sigma_x = GA$
$\sigma_y = GB$

La droite GC, de pente $\rho \frac{\sigma_y}{\sigma_x}$ est la droite des moindres carrés de Y en x. La droite GD, de pente $\frac{1}{\rho} \frac{\sigma_y}{\sigma_x}$ est la droite des moindres carrés de X en y.

$\rho = 1 \Longleftrightarrow$ droites GC et GD confondues: E_G est une ellipse aplatie: X et Y sont liés par une relation <u>fonctionnelle linéaire</u>.

$\rho = 0 \Longleftrightarrow$ ellipse des variances a pour axes ox et oy.
Cette circonstance peut se produire non seulement lorsque X et Y sont liés en probabilité mais également lorsqu'il existe entre eux une liaison fonctionnelle non monotone.

<u>Donc</u>: Lorsqu'on étudie des distributions très générales: ρ est un mauvais indice de corrélation; le rapport de corrélation R est plus significatif.

5.- CAS DES DISTRIBUTIONS A REGRESSION LINEAIRE

Soit $\mathcal{D}_1$ la famille des distributions sur (X,Y) telles que la courbe de régression de Y en x (voir 3) soit une <u>droite</u>. On dit alors que la régression (de Y en X) est <u>linéaire</u>.

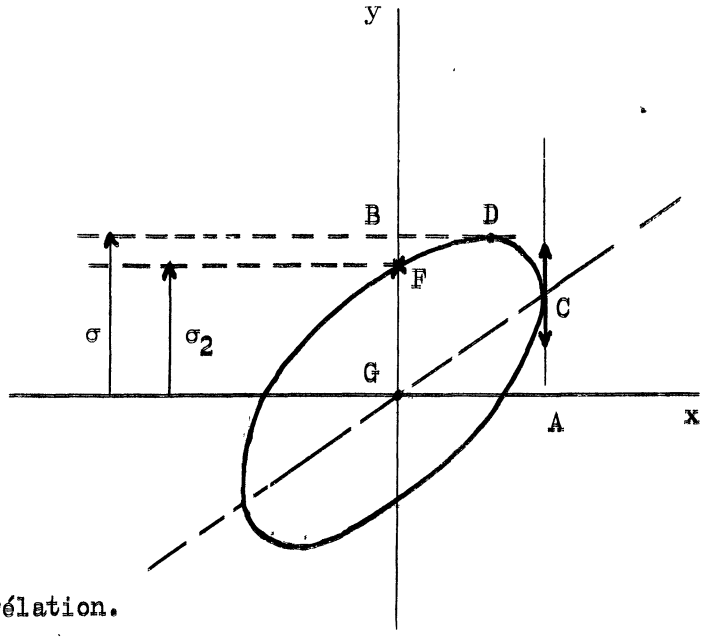
<u>On démontre alors que</u>:
La droite des moindres carrés s'identifie à la droite de régression.

- La variance conditionnelle de Y si x est alors $\sigma_2^2 = \overline{GF}^2$

- Enfin le rapport de régression R s'identifie au coefficient de corrélation:

$$\rho = \frac{\mu_{11}}{\sigma_x \sigma_y} = 1 - \frac{\sigma_2}{\sigma^2}$$

où σ^2 σ_y^2 est la variance totale de Y.
σ_2^2 est la variance conditionnelle de Y.



<u>Conclusion</u>: Au sein de la famille des distributions $\mathcal{D}_1$, le coefficient de corrélation

$$\rho = \frac{\mu_{11}}{\sigma_x \sigma_y}$$

est un indice significatif de la corrélation.

6.- UN MODELE PARTICULIER DE LIAISON: LA DISTRIBUTION DE LAPLACE GAUSS A DEUX DIMENSIONS

La distribution dite "de Laplace-Gauss à deux dimensions", étudiée pour la première fois par BRAVAIS (1846), fut utilisée par F. GALTON; K. PEARSON et après eux par de très nombreux auteurs. Elle présente une structure relativement simple et est susceptible de nombreuses applications. Il importe toutefois, avant de l'adopter comme modèle, de faire preuve d'esprit critique.

Expression de la loi de Laplace-Gauss à deux dimensions: loi élémentaire:

$$(1) \qquad \frac{1}{2n}\,\frac{1}{\sigma_x\,\sigma_y\sqrt{1-\rho^2}}\,e^{-\frac{1}{2}\,\frac{1}{1-\rho^2}\left[\frac{x^2}{\sigma_x^2}-2\,\frac{xy}{\sigma_x\,\sigma_y}+\frac{y^2}{\sigma_y^2}\right]}\,dx\,dy$$

Les courbes réunissant les points d'égale densité sont des ellipses homothétiques et concentriques. Le centre commun à toutes ces ellipses est le centre de la distribution, de coordonnées $m_x = E(X)$ et $m_y = E(X)$. On a choisi ce point pour origine des axes pour écrire la relation (1).

L'une de ces ellipses est l'ellipse des variances (E) dont l'équation est :

$$(2) \qquad \boxed{\frac{x^2}{\sigma_x^2}-2\,\rho\,\frac{xy}{\sigma_x\,\sigma_y}+\frac{y^2}{\sigma_y^2}=1-\rho^2}$$

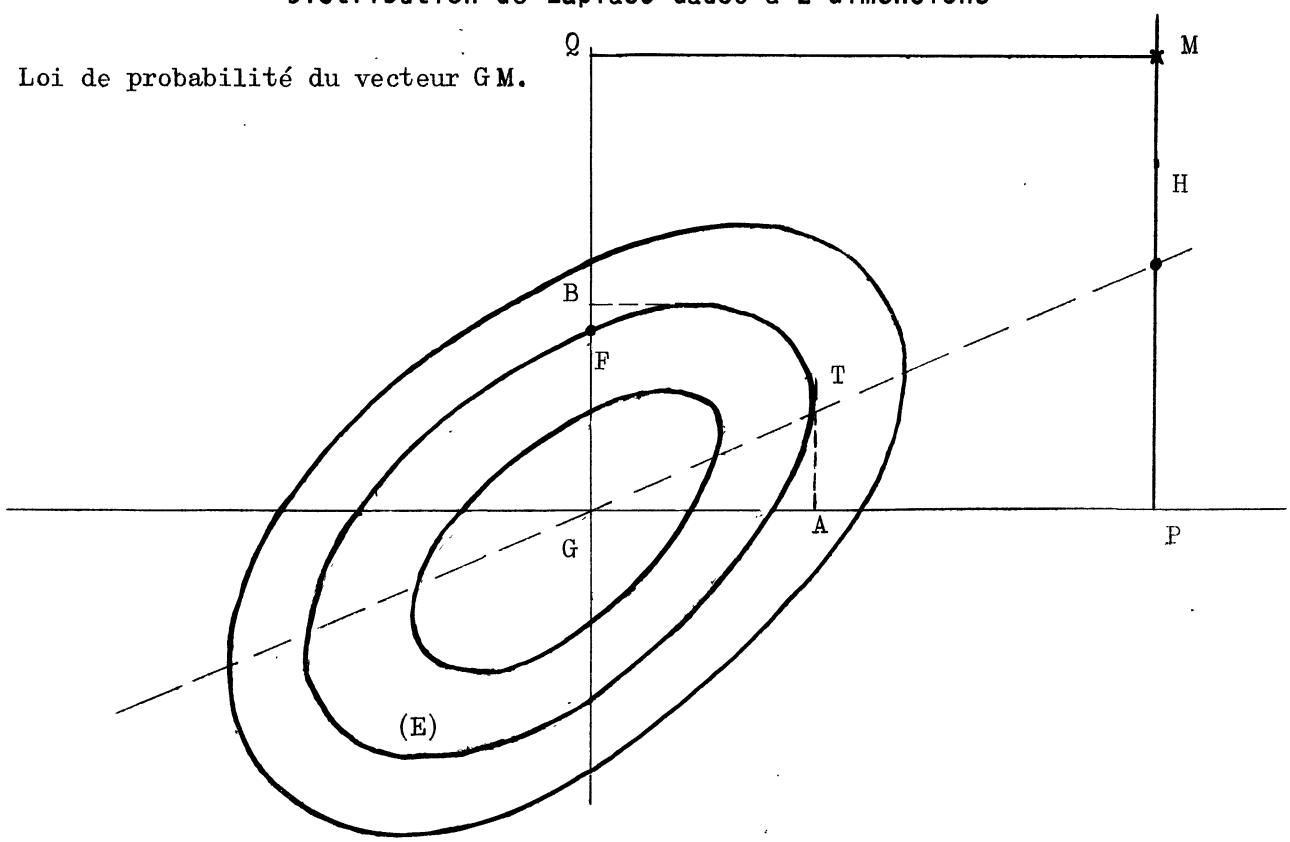
Dans ce modèle: les courbes de régression sont des droites; on dit que les régressions sont linéaires et par conséquent toutes les particularités (5) sont réalisées.

Ces propriétés (5) se rattachent aux moments du 1er et du 2e ordre; et ne font pas état des lois de probabilité des variables étudiées.

La famille des distributions de Laplace-Gauss étudiées maintenant ne contient que 5 paramètres: m_x; m_y; σ_x; σ_y et ρ par exemple. Or la donnée de l'ellipse E des variances définit ces 5 paramètres. Donc dans ce cas, non seulement les moments du 2ème ordre des formes linéaires en (X,Y) sont déterminées, mais les lois de probabilité de ces formes ainsi que les lois des variables liées.

Distribution de Laplace Gauss à 2 dimensions

Loi de probabilité du vecteur G M.



Les trois ellipses sont des courbes joignant les points de même densité (courbes de niveau de la surface figurative des densités). E est l'ellipse des variances.

Le vecteur aléatoire $\overrightarrow{GP}$ obéit à la loi de L.G. (centre G, écart type GA)

Le vecteur aléatoire $\overrightarrow{GQ}$ obéit à la loi de L.G. (centre G, écart type GB)

Si P est donné : $\overrightarrow{PM}$ obéit à la loi de L.G. (centre H, écart type GF).

La droite GT est à la fois: la ligne de régression de Y en x et la droite des moindres carrés de Y en x.

La v.a. X obéit à la loi de Laplace-Gauss de moyenne m_x et d'écart type σ_x. De même pour Y.

Si X = x la v.a. Y obéit à la loi de Laplace-Gauss de moyenne:

$$\overline{y}^{(x)} = m_y + \rho \frac{\sigma_y}{\sigma_x} (x - m_x)$$

son écart type est :

$$\sigma_{y/x} = \sigma_y \sqrt{1 - \rho^2}$$

Cet écart-type est indépendant de x :

il y a homos cédasticité

Il y a même plus: la loi de l'écart de Y (rapporté à la moyenne conditionnelle) c'est-à-dire la loi de

$$Y_1 = Y - y_1^{(x)} \quad \text{où} \quad y_1^{(x)} = E\left[Y \text{ si } X = x\right]$$

est indépendante de x.

On dit alors (définition due à S. BERNSTEIN) que la corrélation entre X et Y est dure.

- - - - - - - - - - - - - -

BIBLIOGRAPHIE

R. FERON "Information - Régression - Corrélation" - thèse - Publication ISUP - Vol. V fasc. 34 - 1956 -

G. DARMOIS - Sur certaines formes de liaisons en probabilité - Colloque CNRS - Le Calcul des Probabilités et ses applications - Lyon - 1948 - Publ. CNRS -