Étude d'un coefficient d'indépendance d'une variable Y par rapport à une variable aléatoire X

J. Guy, J. Bass

ETUDE D'UN COEFFICIENT D'INDEPENDANCE D'UNE VARIABLE ALEATOIRE Y PAR RAPPORT A UNE VARIABLE ALEATOIRE X.

Jean GUY*

et Jean BASS**

Résumé :

Soit (X, Y) un couple de variables aléatoires. Soit $\omega_{ij} = Pr\ (X = x_i$ et $Y = y_j)$. On s'intéresse à la manière dont la loi ω_{ij} se rapproche d'une loi d'indépendance. Pour cela, on pose :

$$\omega_{ij} = \mu\, p_i\, q'_j + (1-\mu)\, u_{ij},$$

où $p_i = \sum_j \omega_{ij} = Pr\ (X = x_i)$ est la loi marginale relative à X, puis on cherche la plus grande valeur possible μ_0 du coefficient μ $(0 \leqslant \mu_0 \leqslant 1)$.

On montre que $\mu_0 = \sum_j \inf_i \dfrac{\omega_{ij}}{p_i}$

Grâce à la définition préalable d'une distance entre matrices stochastiques il apparaît que le rapport $d = \dfrac{1-\mu_0}{\mu_0}$ représente la distance de la matrice A des probabilités conditionnelles :

$$a_{ij} = \frac{\omega_{ij}}{p_i} = Pr\ (Y = y_j / X = x_i)$$

à l'ensemble des matrices stochastiques d'indépendance.

On donne enfin quelques exemples d'application à des problèmes concrets:

* U.E.R. Etudes médicales et biologiques de l'Université René Descartes.
45, rue des Saint-Pères. 75006 - Paris.

** U.E.R. 47 de Mathématiques de l'Université Pierre et Marie Curie
Tour 46, 4, place Jussieu, 75280 Paris Cedex 05

I. Enoncé du problème

Soient X et Y deux variables aléatoires prenant respectivement les valeurs $x_1, x_2, ..., x_m$ et $y_1, y_2, ..., y_n$. Appelons :

ω_{ij} la probabilité de $X = x_i, Y = y_j$

p_i la probabilité de $X = x_i$.

q_j la probabilité de $Y = y_j$

(les probabilités p_i et q_j étant supposées non nulles).

Si $\omega_{ij} = p_i q_j$, X et Y sont indépendantes.

Pour $m \geqslant n$, si $\omega_{ij} = p_i \delta_{f(i),j}$ (f(i) étant une application de l'ensemble des indices i sur l'ensemble des indices j), Y devient une fonction de X.

Dans les autres cas, il y a une dépendance stochastique de Y par rapport à X et il convient de chercher un critère simple pour juger la manière dont la dépendance en question se rapproche de l'indépendance.

Le coefficient de corrélation linéaire ρ entre X et Y donne des indications à ce sujet. S'il y a indépendance, il est nul. Mais s'il est nul, on ne peut rien dire. Il peut même arriver que $\rho = 0$ alors que Y est fonction de X, ainsi que cela a lieu si $Y = X^2$, X étant distribuée suivant une loi normale centrée.

Seul le cas où $\rho = \pm 1$ est précis : Y est alors fonction linéaire de X ; réciproquement, $Y = aX + b$ impose $|\rho| = 1$.

Il est donc naturel de chercher d'autres procédés simples pour tester le degré de dépendance de Y par rapport à X. L'objet de ce travail est d'en suggérer un.

Posons :

$$\omega_{ij} = \mu\, p_i q_j + (1 - \mu)\, u_{ij}. \tag{1}$$

p_i est la probabilité pour que $X = x_i$. μ est un nombre compris entre 0 et 1. u_{ij} a, relativement aux indices, les caractères d'une probabilité : $u_{ij} \geqslant 0$, $\sum_{ij} u_{ij} = 1$. On suppose enfin que q_j est aussi une probabilité. On vérifie que

toutes ces conditions sont compatibles. On a alors :

$$p_i = \sum_j u_{ij} \tag{2}$$

$$q_j = \mu\, q'_j + (1-\mu) \sum_i u_{ij} \tag{3}$$

En général $q'_j \neq q_j$.

Si $\mu = 1$, on a $\omega_{ij} = p_i q'_j = p_i q_j$. Il y a indépendance. Si $\mu = 0$, $\omega_{ij} = u_{ij}$.

Si l'on se donne ω_{ij}, μ n'est pas déterminé. En particulier μ peut prendre la valeur 0. Mais en général μ ne peut pas prendre la valeur 1. Il y a une valeur μ_0 **maximale** compatible avec la donnée de ω_{ij}. Cette valeur représente jusqu'à un certain point la manière dont la loi ω_{ij} donnée se rapproche de l'indépendance. C'est elle qui va servir à tester le degré de dépendance de Y par rapport à X.

Mais, avant de faire le calcul de μ_0, nous pouvons interpréter la relation (1). Introduisons m + 2 variables aléatoires indépendantes

Z prenant les valeurs 1 et 0 avec les probabilités μ et $1 - \mu$,

A prenant n valeurs avec les probabilités q'_j.

B_i prenant n valeurs avec les probabilités u_{ij} (i = 1, 2, ..., m).

On choisit la valeur de X. Pour trouver la valeur de Y, une fois X choisie, on commence par tirer au sort la valeur de Z. Si Z = 1, on fait une expérience pour déterminer A et l'on prend pour valeur de Y la valeur trouvée pour A, quelle que soit la valeur initiale de X. Si Z = 0, on fait une expérience pour déterminer la valeur de la variable B_i dont l'indice correspond à la valeur initiale de X, et l'on prend pour valeur de Y la valeur trouvée pour B_i.

Pratiquement, Z résulte d'une épreuve de Bernoulli (pile ou face généralisé). A est le résultat du tirage dans une urne. Aux B_i correspondent m urnes. On essaie Z. Si Z $\neq$ 1, la valeur de Y résulte du tirage dans la première urne. Si Z = 0, on obtient la valeur de Y en tirant une boule de celle des urnes qui correspond à la valeur initialement choisie pour X.

2. Calcul de la valeur maximale μ_0. Compatibilité

L'équation (1) entraîne :

$q'_j \leqslant \dfrac{\omega_{ij}}{p_i}$ (quel que soit i) et par conséquent :

$$\mu q'_j \leqslant \inf_i \frac{\omega_{ij}}{p_i} \tag{4}$$

Il en résulte que :

$$\mu \leqslant \sum_j \left(\inf_i \frac{\omega_{ij}}{p_i}\right). \tag{5}$$

et la valeur maximale de μ est :

$$\mu_0 = \sum_j \left(\inf_i \frac{\omega_{ij}}{p_i}\right) \tag{6}$$

Les valeurs associées des q'_j sont alors données par :

$$\mu_0 q'_j = \inf_i \frac{\omega_{ij}}{p_i} \tag{7}$$

et celles des u_{ij} par (1). Elles sont déterminées, sauf bien entendu si $\mu_0 = 1$.

Montrons que, pour toute valeur de μ telle que $0 < \mu < \mu_0$, le système (1) permet de calculer des q'_j (non complètement déterminés si $\mu < \mu_0$) et des u_{ij}.

On a :

$$\mu q'_j = \frac{\omega_{ij}}{p_i} - (1-\mu)\frac{u_{ij}}{p_i} \leqslant \frac{\omega_{ij}}{p_i} \leqslant \inf_i \frac{\omega_{ij}}{p_i} \tag{8}$$

Donnons-nous des $q'_j \geqslant 0$ satisfaisant à l'inégalité :

$$\mu q'_j \leqslant \inf_i \frac{\omega_{ij}}{p_i} \tag{9}$$

Nous pouvons en outre demander que $\sum_j q'_j = 1$, puisque :

$$\mu \leqslant \mu_0 = \sum_j \left[\inf_i \frac{\omega_{ij}}{p_i} \right]$$

Les q'_j étant ainsi choisis, l'égalité (1) donne les u_{ij}. Ils sont bien tels que $\sum_{ij} u_{ij} = 1$. En outre, ils sont bien positifs, car

$$\omega_{ij} - \mu\, p_i q'_j \geqslant 0.$$

Le problème a donc en général une infinité de solutions. Il en a une seule si $\mu = \mu_0$. Dans ce cas, on peut préciser le comportement des u_{ij}. Soit i_0 l'indice i pour lequel $\frac{\omega_{ij}}{p_i}$ est minimal. On voit que :

(10) $$u_{i_0 j} = 0.$$

3. Relation entre μ et ρ

On remarque tout d'abord que μ, et en particulier μ_0, ne dépend que des probabilités des variables X, Y, et non des valeurs prises par ces variables. On ne change donc pas μ, ou μ_0, si l'on modifie les valeurs X, Y en conservant les probabilités. Au contraire, le coefficient de corrélation linéaire ρ dépend des valeurs prises par X et Y. Mais, pour comparer ρ et μ_0, on peut choisir ces valeurs comme on le veut. Nous supposerons seulement que :

$$EX = EY = 0.$$

A partir de la relation (1), on obtient :

$$\sigma_1^2 = EX^2 = \sum_{ij} \omega_{ij}\, x_i^2 = \mu\, \sigma_1^2 + (1-\mu) \sum_{ij} u_{ij}\, x_i^2$$

soit encore :

(11) $$\sigma_1^2 = \sum_{ij} u_{ij}\, x_i^2$$

et l'on a pour σ_2^2 :

$$(12) \qquad \sigma_2^2 = \sum_{ij} \omega_{ij}\, y_j^2 = \mu \sum_{ij} q'_j\, y_j^2 + (1-\mu) \sum_{ij} u_{ij}\, y_j^2$$

La définition même de ρ donne :

$$(13) \qquad \rho\, \sigma_1\, \sigma_2 = \sum_{ij} \omega_{ij}\, x_i y_j = 0 + (1-\mu) \sum_{ij} u_{ij}\, x_i\, y_j.$$

Donc :

$$(14) \qquad \rho = \frac{1-\mu}{\sigma_1\, \sigma_2} \sum_{ij} u_{ij}\, x_i\, y_j$$

Posons

$$r = \frac{\sum_{ij} u_{ij}\, x_i\, y_j}{\sqrt{\sum_{ij} u_{ij}\, x_i^2} \cdot \sqrt{\sum_{ij} u_{ij}\, y_j^2}}$$

r n'est pas en général un coefficient de corrélation car, pour la loi u_{ij}, Y n'est pas de valeur moyenne nulle. Mais on a bien entendu :

$$|r| \leqslant 1.$$

On obtient :

$$(15) \qquad \rho = \frac{(1-\mu)\, r}{\sigma_2} \sqrt{\sum_{ij} u_{ij}\, y_j^2}$$

et d'après (12)

$$(16) \qquad \sqrt{1-\mu} \sqrt{\sum_{ij} u_{ij}\, y_j^2} \leqslant \sigma_2.$$

Donc :

$$(17) \qquad \rho \leqslant \sqrt{1-\mu}$$

Cette dernière inégalité, valable pour tout μ, l'est en particulier pour μ_0.

4. Exemple du couple de variables aléatoires à deux valeurs

Supposons $m = n = 2$. On va a priori choisir les deux valeurs possibles de X et de Y pour que $EX = 0$, $EY = 0$. Alors elles sont définies à un facteur près, qui n'a pas d'influence sur les calculs et les résultats.

Nous désignerons par p, $1 - p$ les probabilités des valeurs x_1 et x_2 de X, par q, $1 - q$ celles des valeurs y_1 et y_2 de Y, par λ la probabilité du couple (x_1, y_1). A un facteur près, on a nécessairement :

$$x_1 = 1 - p \qquad x_2 = -p$$

$$y_1 = 1 - q \qquad y_2 = -q$$

ce qui permet de construire le tableau usuel suivant des probabilités des couples de valeurs et des probabilités marginales :

Y \ X	$1 - p$	$-p$	
$1 - q$	λ	$q - \lambda$	q
$-q$	$p - \lambda$	$1 + \lambda - p - q$	$1 - q$
	p	$1 - p$	1

On a :

$$p + q - 1 \leqslant \lambda \leqslant \inf(p, q). \tag{18}$$

Il y a indépendance si $\lambda = pq$.

Il n'est pas en général possible de choisir λ, une fois p et q donnés, pour que Y soit fonction de X. Il faut pour cela que $p = q$. Alors, si $\lambda = p = q$, le tableau ci-dessus devient :

$$\begin{pmatrix} p & 0 \\ 0 & 1 - p \end{pmatrix}$$

On trouve facilement :

$$EX^2 = p(1-p) \qquad EY^2 = q(1-q)$$

$$EXY = \lambda - pq.$$

Le coefficient de corrélation entre X et Y est :

$$\rho = \frac{\lambda - pq}{\sqrt{p(1-p)} \cdot \sqrt{q(1-q)}} \tag{19}$$

Pour calculer μ, on écrit :

$$\begin{aligned} \lambda &= \mu\, pr + (1-\mu)\, u_{11} \\ p - \lambda &= \mu\, p(1-r) + (1-\mu)\, u_{12} \\ q - \lambda &= \mu\,(1-p)\, r + (1-\mu)\, u_{21} \\ 1 + \lambda - p - q &= \mu\,(1-p)(1-r) + (1-\mu)\, u_{22} \end{aligned}$$

On a :

$$\mu r = \inf\left(\frac{\lambda}{p}\ ,\ \frac{q-\lambda}{1-p}\right)\ ,\quad \mu(1-r) = \inf\left(\frac{p-\lambda}{p}\ ,\ \frac{1+\lambda-p-q}{1-p}\right)$$

La valeur de r dépend du signe de $\lambda - pq$. La valeur maximale μ_0 de μ est dans tous les cas égale à $\mu_0 = 1 - \dfrac{|\lambda - pq|}{p(1-p)}$, ce qui s'écrit aussi :

$$1 - \mu_0 = \frac{|\lambda - pq|}{p(1-p)} \tag{20}$$

Il s'agit du μ_0 correspondant aux probabilités conditionnelles de Y une fois X choisi. Si l'on permute X et Y, on trouve un μ_0' tel que :

$$1 - \mu_0' = \frac{|\lambda - pq|}{q(1-q)} \tag{21}$$

On constate que :

(22) $$(1-\mu_0)(1-\mu_0') = \rho^2$$

On vérifie que μ_0 et μ_0' sont compris entre 0 et 1 et que :

(23) $$\rho^2 \leqslant \inf(1-\mu_0, 1-\mu_0')$$

5. **Interprétation du coefficient μ_0 à l'aide de la distance de deux matrices stochastiques.**

Nous poserons $a_{ij} = \frac{\omega_{ij}}{p_i}$. a_{ij} est la probabilité conditionnelle de $Y = y_j$, lorsque $X = x_i$. La matrice $A = (a_{ij})$ est une matrice stochastique :

$$a_{ij} \geqslant 0, \qquad \sum_j a_{ij} = 1.$$

Appelons E_{mn} l'ensemble des matrices stochastiques A à m lignes et n colonnes.

Nous allons maintenant définir une distance d(A, B) entre deux matrices $A, B \in E_{mn}$ par :

(24) $$d(A,B) = \frac{1}{m} \sum_{ij} \left| \frac{a_{ij}}{\mu_{Ao}} - \frac{b_{ij}}{\mu_{Bo}} \right|,$$

μ_{Ao} et μ_{Bo} étant les deux constantes entièrement définies à partir des matrices A et B, suivant :

(25) $$\mu_{Ao} = \sum_j (\inf_i a_{ij}), \qquad \mu_{Bo} = \sum_j (\inf_i b_{ij}).$$

Vérifions tout d'abord que d(A, B) est bien une distance.

a) La relation de définition (24) demeure inchangée par permutation de A et de B, d'où d(A, B) = d(B, A). On vérifie de plus immédiatement que d(A, A) = 0.

b) Une valeur nulle de d(A, B) entraîne l'identité des matrices A et B. Remarquons tout d'abord que $d(A,B) = 0 \Rightarrow \mu_{Ao} \neq 0$ et $\mu_{Bo} \neq 0$ (la distance d(A,B) devient infiniment grande si une et une seulement des deux constantes μ_{Ao} et μ_{Bo} est nulle ; elle n'est plus définie si $\mu_{Ao} = \mu_{Bo} = 0$). Par suite d(A,B) = 0 entraîne :

(26) $\dfrac{a_{ij}}{\mu_{Ao}} = \dfrac{b_{ij}}{\mu_{Bo}}$ $(\forall\, i, j)$ c'est-à-dire $a_{ij} = \dfrac{\mu_{Ao}}{\mu_{Bo}}\, b_{ij}$.

Par sommation sur j de la dernière égalité, il vient :

(27) $1 = \dfrac{\mu_{Ao}}{\mu_{Bo}}$, d'où il résulte que $a_{ij} = b_{ij}$ $(\forall\, i, j)$.

c) L'inégalité triangulaire

(28) $$d(A,C) \leqslant d(A, B) + d(B, C)$$

est également valable. Ecrivons :

$$d(A,B) = d'(A',B'),\ d(A,C) = d'(A',C'),\ d(B,C) = d'(B',C'),$$

où A',B' et C' représentent les matrices (non stochastiques cette fois) définies par :

$$A' = \frac{1}{\mu_{Ao}} A, \qquad B' = \frac{1}{\mu_{Bo}} B, \qquad C' = \frac{1}{\mu_{Co}} C,$$

On a :

(29) $$d'(A', B') = \frac{1}{m} \sum_{ij} |a'_{ij} - b'_{ij}|$$

et des formules analogues. Or (29) définit l'une des distances usuelles entre deux matrices de dimensions (m, n), distance pour laquelle on a bien :

$$d'(A', C') \leqslant d'(A', B') + d'(B', C').$$

Lorsque la matrice B devient une matrice stochastique d'indépendance (les m lignes sont alors identiques), on peut poser $b_{ij} = b_j$ $(\forall i)$. Donc :

(30) $$\mu_{Bo} = \sum_j \inf_i (b_j) = \sum_j b_j = 1.$$

Prenons en particulier pour matrice B la matrice d'indépendance Q' dont les éléments q'_j découlent de ceux de la matrice A par la résolution des équations

(31) $$a_{ij} = \mu_{Ao}\ q'_j + (1 - \mu_{A_o})\ v_{ij}$$

(v_{ij} étant automatiquement les éléments d'une nouvelle matrice stochastique V). On obtient :

(32) $$d(A,Q') = \frac{1}{m} \sum_{ij} \left| \frac{a_{ij}}{\mu_{Ao}} - q'_j \right|$$

$$= \frac{1}{m} \sum_{ij} \left| q'_j + \frac{1-\mu_{Ao}}{\mu_{Ao}} v_{ij} - q'_j \right|$$

$$= \frac{1}{m} \frac{1-\mu_{Ao}}{\mu_{Ao}} \sum_{ij} v_{ij} = \frac{1}{m} \cdot \frac{1-\mu_{Ao}}{\mu_{Ao}} \sum_{i} 1$$

$$= \frac{1-\mu_{Ao}}{\mu_{Ao}}$$

$\frac{1-\mu_{Ao}}{\mu_{Ao}}$ est donc la distance de la matrice A à la matrice d'indépendance qui lui est associée par la formule (31).

Nous voulons montrer que, si b_j représente les éléments d'une matrice stochastique d'indépendance quelconque, on a :

(33) $$\frac{1}{m} \sum_{ij} \left| \frac{a_{ij}}{\mu_{Ao}} - b_j \right| \geqslant \frac{1-\mu_{Ao}}{\mu_{Ao}}$$

Or, par l'hypothèse,

$$a_{ij} = \mu_{Ao}\ q'_j + (1-\mu_{Ao})\ v_{ij}$$

Nous poserons $b_j = q'_j + r_j$.

Le premier membre de (33) s'écrit donc :

(34) $$\frac{1}{m} \sum_{ij} \left| \frac{1-\mu_{Ao}}{\mu_{Ao}} v_{ij} - r_j \right|$$

$1 - \mu_{Ao}$ est différent de 0. Nous savons que certains des v_{ij} sont nuls. Mais ils ne le sont pas tous. Nous allons d'abord supposer que la matrice $B = (b_j)$ (à n lignes identiques) est suffisamment voisine de la matrice $Q' = q'_j$ pour que, lorsque $v_{ij} \neq 0$, on ait :

$$(35) \qquad \frac{1-\mu_{Ao}}{\mu_{Ao}} \, v_{ij} - r_j > 0.$$

Dans (34), il y a donc des termes qui se réduisent à $|r_j|$ et d'autres qui sont représentés par (35), au facteur $\frac{1}{m}$ près. i étant choisi, faisons dans (34) la sommation par rapport à j. Il y a deux sortes de termes :

$$(36) \quad \sum_{j'} \left(\frac{1-\mu_{Ao}}{\mu_{Ao}} \, v_{ij'} - r_{j'} \right) \qquad (j' \text{ tel que } v_{ij'} \neq 0)$$

$$(37) \quad \sum_{j''} |r_{j''}| \qquad (j'' \text{ tel que } v_{ij''} = 0)$$

Or (36) est la somme de :

$$(38) \quad \sum_{j} \frac{1-\mu_{Ao}}{\mu_{Ao}} \qquad \text{(pour toutes les valeurs de j)}$$

et de :

$$(39) \quad -\sum_{j'} r_{j'} \, .$$

Au total, en ajoutant (36) et (37), on obtient :

$$(40) \quad \sum_{j} \frac{1-\mu_{Ao}}{\mu_{Ao}} \, v_{ij} + \sum_{j''} |r_{j''}| - \sum_{j'} r_{j'}$$

Mais $\sum_{j} r_j = 0$. Donc $\sum_{j'} r_{j'} = - \sum_{j''} r_{j''}$ et l'on obtient :

$$(41) \quad \sum_{j} \frac{1-\mu_{Ao}}{\mu_{Ao}} \, v_{ij} + \sum_{j''} |r_{j''}| + \sum_{j''} r_{j''}$$

soit :

$$\frac{1 - \mu_{Ao}}{\mu_{Ao}} + H, \tag{42}$$

où H est une quantité $\geqslant 0$. On somme enfin par rapport à i, on divise par m, et l'on voit que (34) est la somme de $\frac{1 - \mu_{Ao}}{\mu_{Ao}}$ et d'une quantité $\geqslant 0$. Cela prouve l'inégalité (33).

La démonstration prouve que $\frac{1 - \mu_{Ao}}{\mu_{Ao}}$ est un minimum **relatif** pour la distance entre la matrice stochastique A et l'ensemble E des matrices d'indépendance. Mais E est un ensemble convexe. Si donc la distance entre le point A et l'ensemble E admet un minimum relatif au point Q', c'est aussi le minimum absolu. En effet, s'il existait un point B de E tel que :

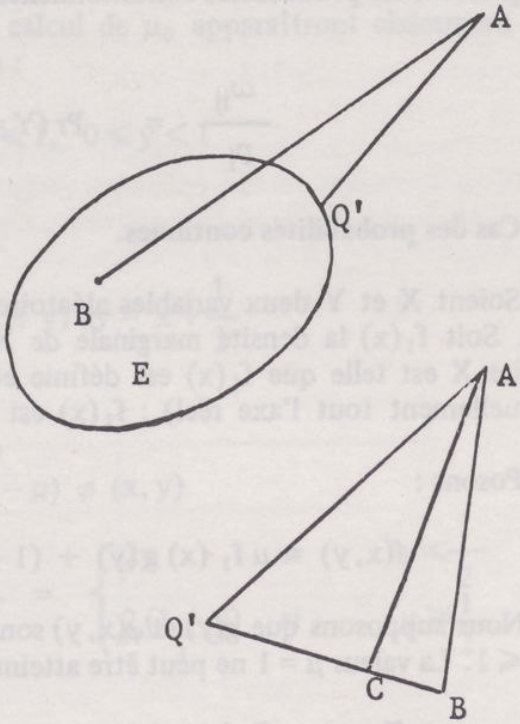

Fig. 1

$$d(A, B) < d(A, Q'),$$

on aurait :

$$d(A, C) < d(A, Q')$$

pour tout C du segment BQ'. Mais $C \in E$ (cf. fig. 1). Il y aurait donc, aussi près qu'on le veut de Q', un point C tel que $d(A, C) < d(A, Q')$ et $d(A, Q')$ ne serait pas un minimum relatif.

On peut conclure de (33) que :

$$d(A, Q') = \frac{1 - \mu_{Ao}}{\mu_{Ao}}$$

est la distance de la matrice stochastique A à l'ensemble des matrices stochastiques d'indépendance. La matrice Q', complètement définie par les équations (31), est la matrice d'indépendance la plus proche de la matrice stochastique A qui représente les probabilités conditionnelles :

$$\frac{\omega_{ij}}{p_i} = \Pr(Y = y_j / X = x_i).$$

6. Cas des probabilités continues.

Soient X et Y deux variables aléatoires ayant une densité de probabilité f(x,y). Soit $f_1(x)$ la densité marginale de X. Nous supposons que la variable aléatoire X est telle que $f_1(x)$ est définie et ne s'annule pas sur un intervalle (éventuellement tout l'axe réel) ; $f_1(x)$ est nulle en dehors de cet intervalle.

Posons :

$$f(x, y) = \mu f_1(x)\, g(y) + (1 - \mu)\, \varphi(x, y) \tag{43}$$

Nous supposons que g(y) et $\varphi(x, y)$ sont des densités de probabilité et que $0 \leqslant \mu \leqslant 1$. La valeur $\mu = 1$ ne peut être atteinte que si :

$$f(x, y) = f_1(x)\, g(y) \tag{44}$$

Il y a alors indépendance et g(y) coïncide avec la densité de probabilité marginale $f_2(y)$ de Y.

Dans le cas général, on a :

$$\mu f_1(x)\, g(y) \leqslant f(x, y) \tag{45}$$

On en déduit :

$$\mu\, g(y) \leqslant \frac{f(x, y)}{f_1(x)} \tag{46}$$

(densité de probabilité conditionnelle de Y pour X = x). Par suite :

$$\mu\, g(y) \leqslant \inf_x \frac{f(x, y)}{f_1(x)} \tag{47}$$

La valeur maximale μ_0 de μ est telle que :

$$\mu_0\, g(y) = \inf_x \frac{f(x,y)}{f_1(x)} \tag{48}$$

On voit que :

$$\mu_0 = \int_{-\infty}^{+\infty} \inf_x \frac{f(x,y)}{f_1(x)}\, dy. \tag{49}$$

Exemple : Les modalités de calcul de μ_0 apparaîtront clairement sur le très simple exemple suivant : Posons :

$f(x,y) = x+y$ pour $0 < x \leqslant 1$, $0 \leqslant y < 1$

$f(x,y) = 0$ ailleurs.

On a $f_1(x) = \int_0^1 (x+y)\, dy = x + \frac{1}{2}$

On pose :

$$x + y = \mu(x + \tfrac{1}{2})\, g(y) + (1-\mu)\, \varphi(x,y)$$

$$g(y)\ \mu_0 = \inf_{0 \leqslant x \leqslant 1} \frac{x+y}{x+\frac{1}{2}} = \begin{cases} 2y & \text{si } y \leqslant \frac{1}{2} \\ \frac{2(1+y)}{3} & \text{si } y \geqslant \frac{1}{2} \end{cases}$$

Donc :

$$\mu_0 = \int_0^{\frac{1}{2}} 2y\, dy + \int_{\frac{1}{2}}^1 \frac{2(1+y)}{3} = \frac{5}{6}$$

7. Exemples d'applications.

Dès qu'une variable sléatoire Y est stochastiquement dépendante d'une autre variable aléatoire X, le coefficient d'indépendance μ_0 défini par (6) et la loi d'indépendance q_j' la plus proche de la loi conditionnelle $\frac{\omega_{ij}}{p_i}$ représenteront des caractéristiques intéressantes pour ce couple de variables. Les quelques exemples simples traités ci-après sont destinés à mieux préciser les informations

liées à la connaissance de μ_0 et de la loi q_j' .

a) Etude de la corrélation existant entre les sexes des jumeaux.

A partir des données statistiques concernant les grossesses gémellaires, il est possible d'estimer les diverses probabilités pour que les deux jumeaux soient, dans l'ordre de leur naissance, garçon-garçon (ω_{GG}), garçon-fille (ω_{GF}), fille-garçon (ω_{FG}) et fille-fille (ω_{FF}). On peut considérer comme valables les valeurs numériques :

$$\omega_{GG} = 0{,}329, \qquad \omega_{GF} = \omega_{FG} = 0{,}179.$$

$$\omega_{FF} = 0{,}313.$$

qui conduisent au tableau suivant représentant la matrice stochastique A (probabilités conditionnelle pour que le second jumeau soit un garçon ou une fille lorsque le sexe du premier jumeau est connu).

1er jumeau \ 2e jumeau	G	F
G	0,648	0,352
F	0,364	0,636

Par suite $\mu_0 = 0{,}364 + 0{,}352 = 0{,}716$ et la distance à la loi d'indépendance la plus proche est

$$d = \frac{1}{\mu_0} - 1 = 0{,}397.$$

Remarquons ici que l'égalité $\omega_{GF} = \omega_{FG}$ entraîne $\rho = 1 - \mu_0 = 0{,}284$ La loi q_j' apparaît ensuite nettement si nous décomposons la matrice A conformément à (31), soit :

$$(50) \qquad A = \begin{pmatrix} 0{,}648 & 0{,}352 \\ 0{,}364 & 0{,}636 \end{pmatrix}$$

$$= \mu_0 \begin{pmatrix} 0{,}508 & 0{,}492 \\ 0{,}508 & 0{,}492 \end{pmatrix} + (1 - \mu_0) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Le complément de μ_0 à l'unité $(1 - \mu_0 = 0{,}284)$ avait déjà été interprété par E. Borel [1]. Ce complément précise le nombre de fois où le sexe du second enfant est complètement déterminé par le sexe du premier (jumeaux «vrais»): l'identité des sexes, due alors à l'identité des génotypes, découle analytiquement de la présence de la matrice unité dans le terme $(1 - \mu_0)$ $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ de la décomposition (50).

b) **Valeurs de μ_0 dans les phénomènes d'hérédité liée au sexe.**

Dans le cas des espèces animales analogues à l'espèce humaine du point de vue du déterminisme sexuel (chromosomes sexuels XX chez la femelle et XY chez le mâle), considérons un gène porté par le chromosome X et susceptible d'exister sous deux états allèles G et g : G (dominant chez les femelles) conduit à un phénotype normal cependant que le génotype g pour le mâle ou gg pour la femelle conduit à un phénotype taré. Si η est la proportion de chromosomes X porteurs de g pour l'ensemble de la population (les croisements étant effectués au hasard) et si α et β sont les proportions de mâles et de femelles, on obtient sans difficultés le tableau suivant (matrice stochastique A) pour les probabilités conditionnelles qu'un sujet soit mâle ou femelle lorsque nous savons que son phénotype est normal (N) ou taré (T).

Phénotype \ Sexe	♂	♀
N	$\dfrac{\alpha(1-\eta)}{1-\alpha\eta-\beta\eta^2}$	$\dfrac{\beta(1-\eta^2)}{1-\alpha\eta-\beta\eta^2}$
T	$\dfrac{\alpha}{\alpha+\beta\eta}$	$\dfrac{\beta\eta}{\alpha+\beta\eta}$

Pour $\alpha = \beta = \dfrac{1}{2}$, nous trouvons :

$$\mu_0 = \frac{1}{2+\eta} + \frac{\eta}{1+\eta} = \frac{1+3\eta+\eta^2}{(2+\eta)(1+\eta)} \tag{51}$$

Les courbes représentant μ_0 et $d = \dfrac{1}{\mu_0} - 1$ en fonction de η sont portées sur la fig. 2. On constate notamment que $0{,}5 < \mu_0 < 0{,}833$, les limites indiquées étant respectivement atteintes de manière asymptotique pour η tendant vers zéro ou vers l'unité. Deux statistiques effectuées séparément sur des sujets

normaux et sur des sujets tarés doivent donc permettre d'atteindre μ_0 : si la valeur trouvée pour μ_0 diffère significativement de l'unité, on pourra conclure à l'existence d'une hérédité liée au sexe et un tel phénomène sera d'autant plus facile à mettre en évidence par l'intermédiaire de μ_0 que la tare sera moins répandue, puisque η très faible conduit à la valeur minimale de μ_0.

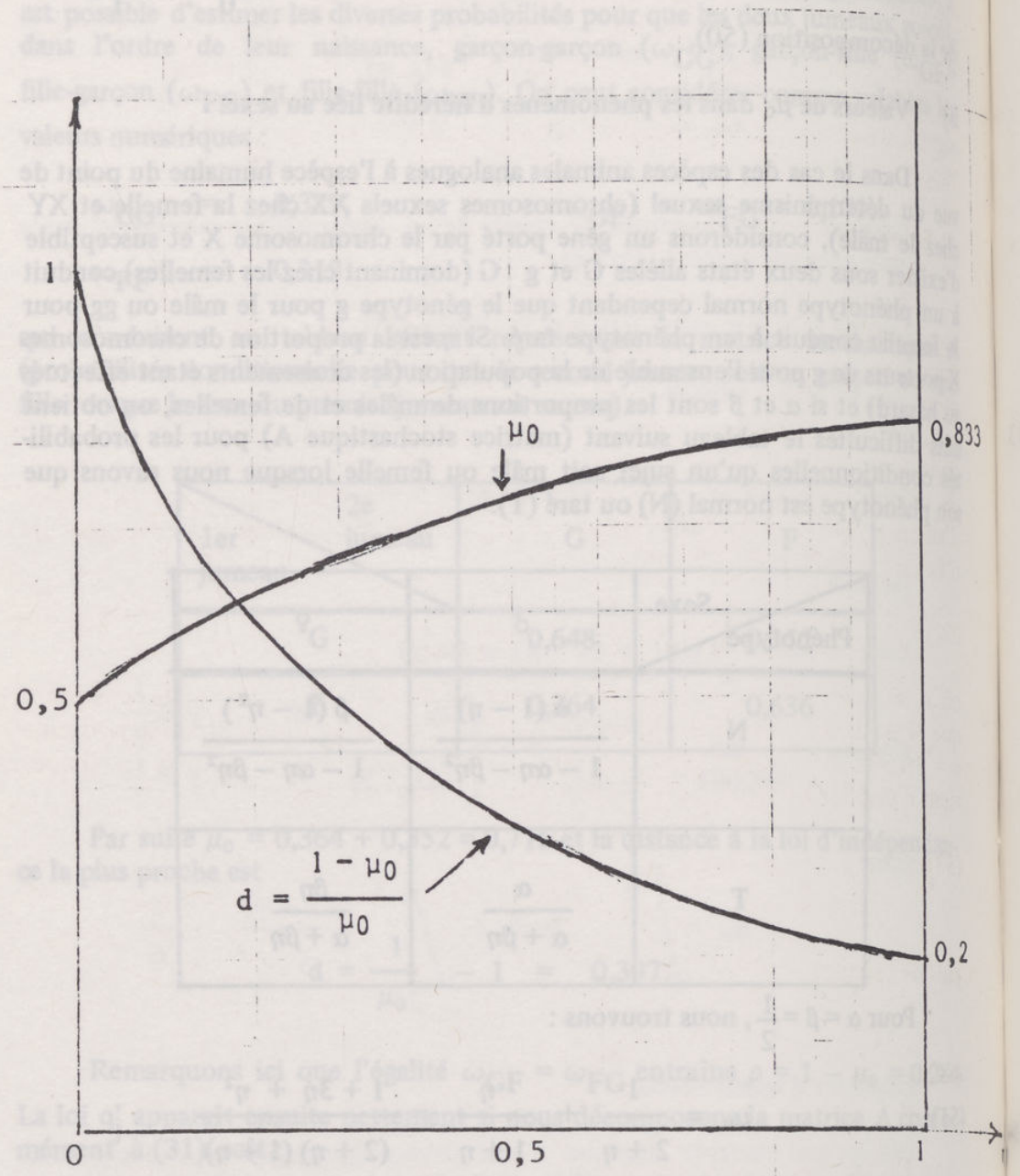

Fig. 2

c) **Possibilité de tester les phénomènes de pléiotropie par un coefficient d'indépendance.**

Toujours en biologie, il y a pléiotropie lorsqu'un gène comportant plusieurs états allèles participe simultanément à la réalisation de plusieurs caractères phénotypiques. Il conviendra de former ici la matrice stochastique A décrivant les probabilités conditionnelles des phénotypes possibles pour le deuxième caractère lorsque les phénotypes relatifs au premier caractère sont connus. Comme il ne peut y avoir indépendance totale entre les deux séries de phénotypes lorsqu'une liaison existe par suite de la pléiotropie, une différence significative de μ_0 (coefficient d'indépendance du deuxième caractère par rapport au premier) vis-à-vis de l'unité signalera un tel phénomène.

d) **Emploi de μ_0 dans certains problèmes de mécanique quantique.**

Notre dernier exemple concernera les distributions électroniques dans les atomes et les molécules. Pour l'état électronique fondamental de l'atome d'hélium, une approximation usuelle est celle du produit simple de deux orbitales atomiques identiques afin de représenter les deux électrons 1 s. C'est là un modèle d'indépendance totale conduisant à $\mu_0 = 1$ lorsque nous étudions les densités de probabilité de présence dans tout l'espace du 2e électron en fonction de la position du premier électron, supposée préalablement fixée. Toutefois, d'autres types de fonctions propres approchées sont possibles, comme celle d'Eckart et Hylleraas [2], soit :

$$\psi(1,2) = N[\exp(-\alpha r_1 - \beta r_2) + \exp(-\beta r_1 - \alpha r_2)], \tag{52}$$

où r_1 et r_2 représentent les distances des électrons 1 et 2 au noyau cependant que N est le facteur usuel de normalisation. L'ajustement des paramètres donne $\alpha = 2{,}15$ et $\beta = 1{,}19$ en unités atomiques de Hartree. Pour le calcul de μ_0, il convient d'évaluer les densités de probabilités conditionnelles :

$$f(2/1) = \frac{\psi^2(1,2)}{\int \psi^2(1,2)\, d\tau_2} \tag{53}$$

puis de calculer :

$$\mu_0 = \int \inf_{(1)} [f(2/1)]\, d\tau_2 \tag{54}$$

On trouve de cette manière $\mu_0 = 0{,}737$ pour la fonction propre approchée (41).

Dans cette application à des variables aléatoires de nature continue (X et Y caractérisent les positions des électrons numérotés 1 et 2), nous vérifions

que les électrons du modèle proposé par Eckart et Hylleraas ne sont plus totalement indépendants bien que μ_0 reste relativement élevé. Notons d'ailleurs que la fonction propre rigoureuse devrait conduire à $\mu_0 = 0$ car $\inf_{(1)} [f(2/1)]$ reste constamment nul pour une bonne fonction propre, la répulsion coulombienne interdisant à deux électrons d'être simultanément présents dans un même élément de volume.

REFERENCES

[1] *BOREL (E.) Eléments de la théorie des probabilités,* p. 155 Hermann édit. (1924)

[2] cf. *PAULING (L.) Introduction to quantum mechanics* p. 224. Mc Graw Hill édit. (1935).

Reçu en Juin 1973.