MATHÉMATIQUES ET SCIENCES HUMAINES

B. MARCHADIER

Dépendance et indépendance de deux aléas numériques images

Mathématiques et sciences humaines, tome 25 (1969), p. 25-34

<http://www.numdam.org/item?id=MSH_1969__25__25_0>

M.S.H. (7e année, no 25, 1969, p. 25-34)

DÉPENDANCE ET INDÉPENDANCE DE DEUX ALÉAS NUMÉRIQUES IMAGES

par

B. MARCHADIER

Cet exposé n'est pas un cours, il cherche :

— à mettre en évidence un certain nombre de notions nécessaires à la compréhension du modèle linéaire, du modèle conditionnel ;

— à définir de manière claire, la signification du coefficient et du rapport de corrélation.

Pour la formulation mathématique et les notations utilisées, on peut se référer au livre de M. Barbut : Mathématiques des Sciences Humaines, *tome II, « Nombres et Mesures », Paris, P.U.F., 1968.*

NOTATIONS.

Rappelons les notations suivantes utilisées dans la suite de cet exposé.

Ω : ensemble des possibilités ;

P : distribution de probabilité sur Ω ;

X : application étagée à nombre fini de valeurs de Ω dans $\mathbf{R}$;

R_X : ensemble des valeurs prises par X ;

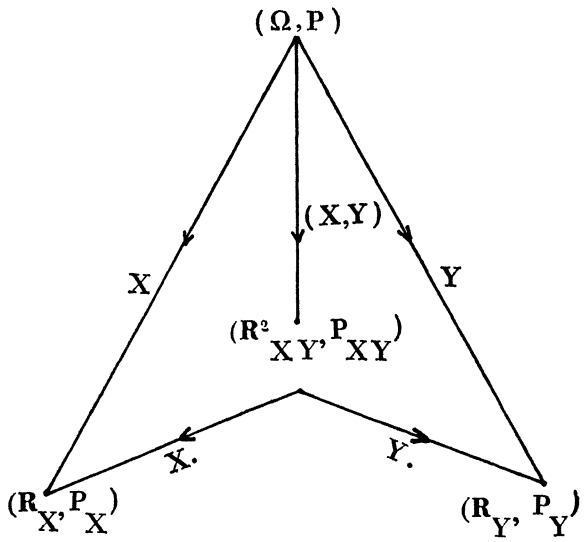
Y : application à nombre fini de valeurs de Ω dans $\mathbf{R}$;

R_Y : l'ensemble des valeurs prises par Y ;

(X, Y) : application de Ω dans $\mathbf{R}^2$ qui a $(w) \to (X(w), Y(w))$;

R_{XY} : l'ensemble des valeurs prises par (X, Y) dans $\mathbf{R}^2$; à l'aléa (Ω, P) correspond par (X, Y) un aléa image dont la loi de probabilité sera notée P_{XY} ; de même lui correspondent par X et Y deux aléas numériques images (R_X, P_X), (R_Y, P_Y).

Le graphique ci-dessous résume le tout.



On voit aisément que X se factorise à travers R^2_{XY}. Si X est l'application projection qui, au couple (x, y) fait correspondre x, on peut écrire :

$$X = X. \circ (X, Y).$$

De même :

$$Y = Y. \circ (X, Y).$$

Les lois P_X et P_Y sont les lois marginales données de la loi du couple P_{XY}.

Remarque.

Dans la suite de ce texte, il nous arrivera de noter par X, l'aléa (R_X, P_X) induit par X s'il n'y a aucune ambiguïté.

De même :

$$Y \simeq (R_Y, P_Y).$$

Rappel.

Théorème de transfert.

Etant donné un aléa numérique (R_X, P_X) et Z une application de R_X dans R induisant l'aléa (R_Z, P_Z), l'espérance de Z peut se calculer indifféremment des deux manières suivantes :

$$\cdot\ E(Z) = \Sigma\ z_i\ P_Z(z_i)$$
$$\cdot\ E(Z) = \Sigma\ Z(x_i)\ P_X(x_i).$$

ALÉAS INDÉPENDANTS.

Indépendance de deux ensembles dans Ω.

C'est à propos d'événements que s'introduit pour la première fois la notion d'indépendance.

$$A \text{ et } B \text{ indépendants} \Leftrightarrow P(A \cap B) = P(A)\ P(B).$$

Indépendance de deux partitions.

Une remarque immédiate :

Si A et B sont indépendants, $\overline{A}$ et $\overline{B}$ aussi, ainsi que $\overline{A}$ et B, A et $\overline{B}$.

En effet :

$$\begin{aligned} P(\overline{A}) \times P(B) &= (1 - P(A))\, P(B) \\ &= P(B) - P(A)\, P(B) \\ &= P(B) - P(A \cap B) \\ &= P(\overline{A} \cap B). \end{aligned}$$

Nous dirons :

— les deux partitions (A, $\overline{A}$) et (B, $\overline{B}$) sont indépendantes l'une de l'autre.

La généralisation est facile et conduit à la définition suivante de l'indépendance de deux partitions A et B quelconques de Ω.

$$A = (A_1, \ldots, A_n)$$
$$B = (B_1, \ldots, B_p)$$

$$A \text{ et } B \text{ indépendantes} \Leftrightarrow \begin{cases} \forall i = 1, \ldots n \\ \forall j = 1, \ldots p \end{cases} P(A_i \cap B_j) = P(A_i) \times P(B_j)$$

Indépendances de deux aléas.

Deux aléas $(\mathbf{R}_X, P_X)$, $(\mathbf{R}_Y, P_Y)$ seront indépendants si les partitions A et B induites par X et Y sur Ω sont indépendantes.

$$X, Y \text{ indépendants} \Leftrightarrow A \text{ et } B \text{ indépendantes.}$$

Ceci se traduit de la manière suivante sur l'aléa image :

Si : $\mathbf{R}_X = \{x_1, \ldots, x_n\}$,
$\mathbf{R}_Y = \{y_1, \ldots, y_p\}$.
Si : $A = (A_1, \ldots, A_n)$ est la partition induite par X, donc telle que :
$X(w) = x_i$ si $w \in A_i$
$B = (B_1, \ldots, B_p)$ la partition induite par Y,

donc telle que :

$$Y(w) = y_j \text{ si } w \in B_j$$

alors :

$$\begin{aligned} P_{XY}\{x_i, y_j\} &= P(A_i \cap B_j) \\ &= P(A_i) \times P(B_j) \\ &= P_X\{x_i\} \times P_Y\{y_j\}. \end{aligned}$$

D'où la définition équivalente de l'indépendance de deux aléas numériques :

$$X, Y \text{ indépendants} \Leftrightarrow P_{XY}(x_i, y_j) = P_X(x_i)\, P_Y(y_j).$$

L'information apportée par la loi du couple se sépare donc en deux parties, l'une concernant $(\mathbf{R}_X, P_X)$, l'autre $(\mathbf{R}_Y, P_Y)$.

Si Y seul nous intéresse, nous pourrons étudier $(\mathbf{R}_Y, P_Y)$ sans considérer $(\mathbf{R}_X, P_X)$, *indépendamment* de $(\mathbf{R}_X, P_X)$.

ALÉAS DÉPENDANTS EN PROBABILITÉ.

Nous supposons maintenant que les deux partitions A et B induites par X et Y sur Ω ne sont plus indépendantes.

L'égalité :

$$P(A_i \cap B_j) = P(A_i) \times P(B_j)$$

n'est plus valable.

Par contre, nous pouvons écrire :

$$P(A_i \cap B_j) = P^{B_j}(A_i) \times P(B_j)$$

ou :

$$= P^{A_i}(B_j) \times P(A_i)$$

et donc :

$$P_{XY}(x_i, y_j) = P_X^{y_j}(x_i) \times P_Y(y_j)$$
$$= P_Y^{x_i}(y_j) \times P_X(x_i)$$

Nous voyons que nous pouvons redéfinir la loi du couple si nous connaissons la loi marginale de X (resp. Y) et les probabilités conditionnelles de Y (resp. de X) pour toutes les valeurs possibles de X (resp. Y). Il est donc possible de remplacer l'étude de la loi du couple par la connaissance :

— d'un aléa marginal, par exemple X ;

— d'un certain nombre d'aléas conditionnels, par exemple les $\{(\mathbf{R}_Y, P_Y^{x_i}, i = 1, \ldots, n\}$.

On peut encore dire, par la connaissance :

— de l'aléa marginal X ;

— de la liaison entre X et Y.

ÉTUDE DE LA LIAISON ENTRE X ET Y.

L'étude des n lois conditionnelles serait fastidieuse. Nous allons schématiser la liaison entre X et Y en résumant ces lois par les espérances et les variances des aléas correspondants.

A l'aide des *espérances*, nous aurons une représentation de la *liaison*, ce sera le graphe de la régression (en abscisse les x_i, en ordonnée les espérances correspondantes).

A l'aide des *variances*, nous aurons une idée de la *force de la liaison.*

Les propriétés de l'espérance et de la variance conditionnelle que nous étudierons dans les paragraphes suivants, vont nous permettre de construire un modèle de liaison entre X et Y.

Espérance conditionnelle – Variance conditionnelle.

Pour chaque valeur $X = x_i$, nous remplaçons donc la distribution conditionnelle par son espérance mathématique et par sa variance.

A chaque valeur x_i, nous associons donc deux valeurs : z_i, l'espérance mathématique et t_i la variance.

Nous définissons ainsi deux applications de $\mathbf{R}_X$ dans $\mathbf{R}$ que nous noterons :

— l'une $Z = E^X Y$,

— l'autre $T = V^X Y = E^X (Y - E^X Y)^2$,

Z étant donc la régression de Y en X.

L'expression littérale des z_i est la suivante :

$$z_i = \sum_{j=1}^{p} y_j \, P_Y^{x_i}(y_j) = Z(x_i).$$

Celle des t_i :

$$t_i = \sum_{j=1}^{p} (y_j - Z(x_i))^2 \, P_Y^{x_i}(y_j) = T(x_i).$$

Propriétés de l'espérance conditionnelle.

L'application $Z = E^X Y$ détermine un nouvel aléa numérique :

$$\begin{array}{c} (R_X, P_X) \\ \Big\downarrow \; Z = E^X Y \\ (R_Z, P_Z) \end{array}$$

— R_Z étant l'ensemble des valeurs possibles de Z ;

— P_Z la distribution de probabilité associée définie par :

$$P_Z(z_i) \quad = \quad P\{x, Z(x) = z_i\}$$

Propriété remarquable de $Z = E^X Y$.

L'espérance de Z égale celle de Y : $E(E^X Y) = E(Y)$.

Montrons le :

$$\begin{aligned} E(Z) &= \Sigma \, z_i \, P_Z(z_i) \\ &= \sum_{i=1}^{n} Z(x_i) \, P_X(x_i) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{p} y_j \, P_Y^{x_i}(y_j) \, P_X(x_i) \\ &= \sum_{j=1}^{p} y_j \, P_Y(y_j) \\ &= E(Y). \end{aligned}$$

$$\boxed{E(E^X Y) = E(Y)}$$

Propriétés de la variance.

L'application $T = V^X Y$ détermine, elle aussi, un nouvel aléa numérique.

$$\begin{array}{c} (R_X, P_X) \\ \Big\downarrow \; T = V^X Y \\ (R_T, P_T) \end{array}$$

$$T = V^X Y = E^X (Y - E^X Y)^2 = E^X (Y^2) - E^X (Y)^2.$$

Propriété remarquable de la variance conditionnelle.

La variance de Y se sépare en la somme de deux termes :

$$V(Y) = E(V^X Y) + V(E^X Y)$$

$E(V^X Y)$ = Espérance de la variance conditionnelle ;

$V(E^X Y)$ = Variance de l'espérance conditionnelle.

Nous avons :

$$E(Y - E^X Y) = E(Y) - E(E^X Y) = E(Y) - E(Y) = 0$$

$$E(E^X Y - E(Y)) = E(E^X Y) - E(E(Y)) = E(Y) - E(Y) = 0$$

$$E(Y - E^X Y)(E^X Y - E(Y)) = 0,$$

ce qui exprime simplement que la covariance de $(Y - E^X Y)$ et $(E^X Y)$ est nulle. Nous allons démontrer de façon détaillée, cette dernière égalité.

$$E(Y - E^X Y)(E^X Y - E(Y)) = E(YE^X Y) - E(E^X Y)^2 - E(Y \times E(E^X Y)) - (E(Y))^2$$

$$= E(YE^X Y) - E(E^X Y)^2 \text{ car } E^X Y = E(Y).$$

Or :

$$E(YE^X Y) = \sum_{ij} y_j Z(x_i) P_{XY}(x_i, y_j) \quad \text{(théorème de transfert) ;}$$

$$= \Sigma\, y_j Z(x_i) P_Y^{x_i}(y_j) \times P_X(x_i)$$

$$= \sum_i \left(\sum_j y_j P_Y^{x_i}(y_j) \right) Z(x_i) P_X x_i$$

$$= \sum_i (Z(x_i))^2 P_X(x_i) = \sum_i z_i^2 P_Z(z_i) \quad \text{(théorème de transfert)}$$

$$= E(E^X Y)^2.$$

La covariance de $Y - E^X Y$ et $E^X Y$ est donc bien nulle.

$$E(V^X Y) = E(Y - E^X Y)^2$$

En effet :

$$V^X Y = T$$

$$E(V^X Y) = \Sigma\, t_l P_T(t_l)$$

$$E(V^X Y) = \sum_{i=1}^{n} T(x_i) P_X(x_i) \quad \text{(théorème de transfert)}$$

$$= \sum_{i=1}^{n} \left(\sum_{j=1}^{p} (y_j - Z(x_i))^2 P_Y^{\{x_i\}}(y_j) \right) P_X(x_i)$$

$$= \sum_{ij} (y_j - Z(x_i))^2 P_{XY}(x_i, y_i)$$

$$= E(Y - E^X Y)^2.$$

Il nous est maintenant possible d'écrire :

$$\begin{aligned} V(Y) &= E(Y - E(Y))^2 \\ &= E((Y - E^X Y) + (E^X Y - E(Y)))^2 \\ &= E(Y - E^X Y)^2 + E(E^X Y - E(Y))^2 \\ &= E(V^X Y) + E(E^X Y - E(Y))^2 \\ &= E(V^X Y) + V(E^X Y). \end{aligned}$$

Modèle de régression conditionnelle.

Les propriétés de l'espérance conditionnelle et de la variance conditionnelle, nous conduisent à poser le modèle suivant :

$$\boxed{Y = E^X Y + e}$$

où e est un aléa numérique indépendant de X et d'espérance nulle.

Propriétés du modèle.

— $E^X Y$ exprime la liaison entre X et Y.

— La variance de Y est complètement reconstituée.

En effet :

$$E(Y - E^X Y) = 0$$

d'où :

$$V(e) = E(Y - E^X Y)^2 = E(V^X Y)$$

nous l'avons déjà vu.

Nous savons aussi que $Y - E^X Y$ et $E^X Y$ sont de covariance nulle.

Par suite :

$$V(E^X Y + e) = E(V^X Y) + V(E^X Y) = V(Y).$$

Rapport de corrélation.

$$V(Y) = V(E^X Y) + V(e) = V(E^X Y) + E(V^X Y)$$

Divisons par V (Y) :

$$1 - \frac{V(E^X Y)}{V(Y)} = \frac{E(V^X Y)}{V(Y)}$$

— La liaison est forte, nous l'avons vu, si les variances conditionnelles sont petites, donc si le second membre est petit.

— La liaison est forte si la part de variance de Y expliquée par la liaison est grande donc si :

$$\frac{V(E^X Y)}{V(Y)}$$

se rapproche de 1.

Inversement.

— La liaison est faible si la part de variance de Y expliquée par e est grande, si le second membre est grand et se rapproche de 1.

— Elle est faible si la part de variance de Y expliquée par la liaison est faible, donc si :

$$\frac{V(E^X Y)}{V(Y)} \text{ tend vers } 0.$$

Nous utiliserons le coefficient η^2, dénommé rapport de corrélation pour juger la force de la liaison en écrivant :

$$\eta^2 = \frac{V(E^X Y)}{V(Y)}$$

$$1 - \eta^2 = \frac{E(V^X Y)}{V(Y)}.$$

Remarque sur le rapport de corrélation.

— Supposons X, Y indépendants, alors :

$$E^X Y = E(Y) \quad \text{et} \quad V(E^X Y) = 0.$$

donc :

$$\eta^2 = 0.$$

— Supposons X, Y liés fonctionnellement : $Y = f(X)$, alors :

$$E^X Y = Y = f(X).$$

Les distributions conditionnelles sont dégénérées.

$$V(e) = 0$$

et

$$\eta^2 = 1.$$

Modèle de régression linéaire.

Nous allons faire une hypothèse restrictive (très restrictive) sur la forme de la courbe représentative de la régression. Nous supposerons que c'est une droite, que l'on peut écrire :

$$E^X Y = a X + b.$$

Le modèle de régression conditionnelle devient alors :

$$\boxed{Y = a X + b + e} \qquad = Y'_X + e.$$

Ce modèle possède les mêmes propriétés que le précédent, compte tenu de l'hypothèse faite.

Coefficient de corrélation.

Pour mesurer la force de la liaison entre X et Y, nous utiliserons le même indice que précédemment :

$$\eta^2 = \frac{V(E^X Y)}{V(Y)}$$

η est alors égal au coefficient de corrélation linéaire ρ. Nous allons voir qu'on peut l'écrire sous une forme qui fait intervenir la covariance de X et Y, la variance de X et la variance de Y.

$$\text{Cov}(X, Y) = E(Y - E(Y))(X - E(X)) = \frac{1}{a} E(Y - E(Y))(aX + b - (aE(X) + b))$$

$$= \frac{1}{a} E(Y - E(Y))(E^X Y - E(Y))$$

$$= \frac{1}{a} E(Y - E^X Y)(E^X Y - E(Y)) + \frac{1}{a} E(E^X Y - E(Y))^2$$

$$= \frac{1}{a} E(E^X Y - E(Y))^2 \text{ car la covariance entre } Y - E^X Y \text{ et } E^X Y \text{ est nulle}$$

$$= \frac{1}{a} V(E^X Y)$$

mais ceci peut s'écrire encore :

$$= \frac{1}{a} E(aX + b - (aE(X) + b))^2$$

$$= \frac{1}{a} E(a(X - E(X)))^2$$

$$= a\, V(X).$$

Donc :

$$\rho^2 = \frac{V(E^X Y)}{V(Y)}$$

$$= \frac{a \,\text{cov}(X, Y)}{V(Y)}$$

$$= \frac{\text{cov}^2(X, Y)}{V(X)\, V(Y)}$$

$$\boxed{\rho = \frac{\text{cov}(X, Y)}{\sqrt{V(X)\,(V(Y)}}}$$

Remarque.

Le signe de ρ nous donnera le sens de la liaison (c'est-à-dire le signe de a) car :

$$\text{cov}(X, Y) = a\, V(X).$$

Comparaison du coefficient de corrélation linéaire et du rapport de corrélation.

— Si la liaison est linéaire, les deux coefficients coïncident évidemment.

$$\rho^2 = \eta^2 = \begin{cases} 1 & \text{liaison affine fonctionnelle} \\ 0 & \text{variables indépendantes.} \end{cases}$$

$0 < \rho^2 = \eta^2 < 1$ liaison linéaire plus résidu aléatoire

— Si la liaison n'est pas linéaire, ou pas parfaitement linéaire, on a toujours :

$$\rho^2 < \eta^2$$

$\rho^2 < \eta^2 = 1$ liaison fonctionnelle non linéaire

$0 < \rho^2 < \eta^2 < 1$ liaison non linéaire.

Montrons que ρ^2 est toujours plus petit que η^2 :

$$\rho^2 = \frac{(\text{cov}\,(X, Y))^2}{V\,(X)\,V\,(Y)}$$

$$\eta^2 = \frac{V\,(E^X\,Y)}{V\,(Y)}$$

Il suffit donc de montrer que :

$$V\,(E^X\,Y) > \frac{(\text{cov}\,(X, Y))^2}{V\,(X)}$$

D'après l'inégalité de Schwarz :

$$(\text{cov}\,(X, Y))^2 \leqq V\,(X) \cdot V\,(Y)$$

or :

$$V\,(E^X\,Y) = V\,(Y) - E\,(V^X\,Y) \leqslant V\,(Y)$$

donc :

$$(\text{cov}\,(X, Y))^2 \leqslant V\,(X)\,V\,(E^X\,Y)$$

et

$$V\,(E^X\,Y) \geqq \frac{(\text{cov}\,(X, Y))^2}{V\,(X)}$$

$$\boxed{\rho^2 < \eta^2}$$