

Liaisons entre deux variables quantitatives *ou* entre une variable quantitative et une variable qualitative

Jean-Claude Régnier

Professeur des Universités
Université Lumière Lyon2

Comment peut-on explorer les liaisons entre deux variables statistiques ?

Lorsque nous sommes conduit à étudier deux variables statistiques X et Y sur une population P, il est parfois intéressant de rechercher si ces deux variables sont indépendantes ou dépendantes. Nous avons déjà traité le cas où les deux variables sont qualitatives¹. Considérons d'abord celui où celles-ci sont quantitatives puis celui où l'une est quantitative et l'autre est qualitative. Pour exposer les notions mises en jeu nous allons utiliser un exemple simulé. Pour cela nous rapportons le tableau statistique des résultats de quatre variables sur une population finie de 25 unités.

individus	V1	V2	V3	V4	individus	V1	V2	V3	V4
w1	15	0	0	C	w14	18	9	9	B
w2	15	0	9	C	w15	18	9	9	B
w3	15	0	9	C	w16	12	6	0	A
w4	18	9	9	A	w17	15	0	6	C
w5	18	9	0	B	w18	12	6	6	A
w6	18	9	0	B	w19	15	0	0	C
w7	15	0	9	C	w20	18	9	6	A
w8	18	9	0	B	w21	15	0	0	C
w9	15	0	0	C	w22	15	0	9	C
w10	12	6	0	A	w23	15	0	6	C
w11	12	6	9	A	w24	18	9	0	B
w12	18	9	6	B	w25	18	9	6	B
w13	12	6	9	B					

De ce tableau des séries statistiques nous pouvons extraire les informations utiles pour déterminer les distributions conjointes de fréquence des couples de variables (V1,V2), (V1,V3) et (V1,V4) et les distributions marginales de fréquence de chaque variables.

¹ Régnier, J.-C., (1996) *Méthodes quantitatives et statistique* cours photocopié de licence & maîtrise de sciences de l'éducation, univ lyon2 , tome 1 pp 55-60

	V2			V3			V4			
V1	0	6	9	0	6	9	A	B	C	effectifs
12	0	5	0	2	1	2	4	1	0	5
15	10	0	0	4	2	4	0	0	10	10
18	0	0	10	4	2	4	2	8	0	10
effectifs	10	5	10	10	5	10	6	9	10	25

De là nous calculons les paramètres usuels de V1, V2 et V3 :

variables	moyenne	variance	covariance
V1	$m(V1) = 15,6$	$var(V1) = 5,04$	
V2	$m(V2) = 4,8$	$var(V2) = 16,56$	$cov(V1, V2) = 4,32$
V3	$m(V3) = 4,8$	$var(V3) = 16,56$	$cov(V1, V3) = 0$

La covariance est définie par la relation suivante :

$$\text{cov}(X, Y) = m[(X - m(X))(Y - m(Y))] = m(XY) - m(X)m(Y)$$

Nous pouvons aussi faire apparaître les fréquences conditionnelles par les profils lignes et les profils colonnes :

	V2			V3			V4			
V1	0	6	9	0	6	9	A	B	C	
12	0	1	0	0,4	0,2	0,4	0,8	0,2	0	1
15	1	0	0	0,4	0,2	0,4	0	0	1	1
18	0	0	1	0,4	0,2	0,4	0,2	0,8	0	1
profil moyen	0,4	0,2	0,4	0,4	0,2	0,4	0,24	0,36	0,4	1

	V2			V3			V4			profil moyen
V1	0	6	9	0	6	9	A	B	C	
12	0	1	0	0,2	0,2	0,2	2/3	1/9	0	0,2
15	1	0	0	0,4	0,4	0,4	0	0	1	0,4
18	0	0	1	0,4	0,4	0,4	1/3	8/9	0	0,4
	1	1	1	1	1	1	1	1	1	1

Rappelons que l'indépendance statistique de deux variables X et Y est définie par la relation suivante : $\text{Pr}[X \text{ sachant } Y] = \text{Pr}[X]$ traduisant ainsi que la fréquence d'apparition d'un résultat quelconque de X ne dépend d'aucune condition sur un résultat quelconque de Y. D'où la caractéristique algébrique de l'indépendance statistique de X et de Y :

$$\text{Pr}[X \text{ et } Y] = \text{Pr}[X] \text{Pr}[Y]$$

Par négation, s'il existe au moins un résultat de X pour lequel $\text{Pr}[X \text{ sachant } Y] \neq \text{Pr}[X]$, c'est dire tel que $(\text{Pr}[X \text{ et } Y] \neq \text{Pr}[X] \text{Pr}[Y])$, on dit alors que les deux variables sont dépendantes statistiquement. Naturellement ce lien est à analyser de très près car il n'est pas nécessairement causal, cependant si ce lien est explicitable il peut permettre de prévoir le

résultat d'une variable connaissant l'autre. C'est le cas si nous pouvons trouver une relation fonctionnelle mathématique explicite à partir de laquelle des calculs sont possibles.

Si nous observons les données de notre exemple, nous pouvons remarquer que V1 et V3 sont deux variables statistiquement indépendantes alors que V1 et V2 ainsi V1 et V4 sont statistiquement dépendantes. L'indépendance de V1 et V3 justifie la nullité de la covariance $cov(V1,V3)$. Notons que la covariance de deux variables peut cependant être aussi nulle, même si les deux variables sont dépendantes². En revanche la non nullité de la covariance implique l'existence d'une dépendance statistique entre les deux variables.

Coefficient r de corrélation linéaire de Bravais-Pearson et coefficient empirique R_{BP} .

Lorsque nous avons affaire à un couple de variables quantitatives, une première approche de l'étude de la liaison consiste à construire une représentation graphique géométrique dans laquelle les points ont pour coordonnées les couples de résultats (x,y) . La forme du **nuage statistique** suggère des pistes pour donner un sens à une liaison possible entre X et Y.

On définit le coefficient ρ de corrélation linéaire de Bravais-Pearson par la relation algébrique³ suivante : $\rho = \frac{cov(X,Y)}{\sigma(X)\sigma(Y)}$ ou encore $\rho^2 = \frac{cov^2(X,Y)}{V(X)V(Y)}$.

On montre que $-1 \leq \rho \leq 1$. Ce coefficient est nul si les deux variables sont indépendantes. En revanche si la nullité de ρ exclut l'existence d'une **relation linéaire** entre X et Y, elle n'exclut pas l'existence d'autres relations et même des relations fonctionnelles⁴.

Par ailleurs $\rho = \pm 1$ s'il y a une relation linéaire entre les deux variables X et Y. Cette propriété serait suggérée par la forme rectiligne du nuage.

$$\text{Dans l'exemple } \rho(V1,V2) = \frac{4,32}{\sqrt{5,04}\sqrt{16,56}} = 0,4728 \text{ et } \rho(V1,V3) = 0$$

Le problème est que nous ne possédons la plupart du temps que des données d'échantillon. Il convient alors de définir une *statistique* permettant d'estimer la valeur ρ inconnue ou de prendre une décision dans un test d'hypothèse relatif à ce coefficient.

Considérons alors un n-échantillon (X_i, Y_i) du couple (X,Y) . On définit le coefficient de corrélation linéaire empirique de Bravais-Pearson de la manière suivante :

² Régnier, J-C, (1997) *Indépendance de deux variables et covariance* article photocopié pour cours de licence & maîtrise de sciences de l'éducation, univ Lyon2 , 6 p.

³ Ce coefficient peut être interprété géométrique comme le cosinus d'un angle dans un espace bien choisi, la variance comme une norme et la covariance comme un produit scalaire.

⁴ *ibidem*

$$R_{BP} = \frac{\text{COV}(X_n, Y_n)}{S_{X_n} S_{Y_n}}$$

qui est obtenu à partir de **moyenne empirique** définie par $X_n = \frac{1}{n} \sum_{i=1}^n X_i$ et $Y_n = \frac{1}{n} \sum_{i=1}^n Y_i$

variance empirique est définie par $S_{X_n}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - X_n)^2$ et $S_{Y_n}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - Y_n)^2$

covariance empirique est définie par $\text{COV}(X_n, Y_n) = \frac{1}{n} \sum_{i=1}^n (X_i - X)(Y_i - Y)$

Lorsque le couple (X,Y) est un couple binormal⁵, les paramètres de cette statistique sont connus et admettent les valeurs approximatives :

$$\begin{array}{ll} E(R_{BP}) & \rho - \frac{\rho(1-\rho^2)}{2n} \\ \gamma_1 & \frac{-6\rho}{\sqrt{n}} \end{array} \quad \begin{array}{l} V(R_{BP}) \\ \gamma_2 \end{array} \quad \begin{array}{l} \frac{(1-\rho^2)^2}{n-1} \left(1 + \frac{11\rho^2}{2n} \right) \\ \frac{6(12\rho^2-1)}{n} \end{array}$$

Caractère significatif du coefficient de corrélation de Bravais-Pearson :

Avant toute étude plus approfondie, il convient de réaliser une représentation graphique du nuage statistique des n points de coordonnées (x_i, y_i) La forme de ce nuage orientera l'analyse.

Supposons que les n observations proviennent d'une population dans laquelle les deux variables X et Y sont indépendantes. Dans ce cas la valeur réelle du coefficient de corrélation est $\rho = 0$. On peut alors utiliser la distribution de probabilité de la statistique R_{BP} correspondant à cet échantillonnage. Il est établi que si $\rho = 0$ et si le couple (X,Y) est un couple de variables de Laplace -Gauss, l'espérance de R_{BP} vaut $E(R_{BP}) = 0$ et la variance

$$V(R_{BP}) = \frac{1}{n-1} . \text{ De plus, la distribution de probabilité de la variable transformée } \frac{R_{BP} \sqrt{n-2}}{\sqrt{1-R_{BP}^2}}$$

est celle de la variable de Student de ddl = n-2

Dans le cas général où ρ est quelconque dans [-1 ; +1] , on peut utiliser la transformée de Fisher :

⁵ variable de Laplace Gauss à deux dimensions dont la densité est :

$$f(x,y) = \frac{1}{2\delta\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y-\mu_2}{\sigma_2}\right)^2 - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right]\right\}$$

$$Z = \frac{1}{2} \ln \left(\frac{1+R_{BP}}{1-R_{BP}} \right) \xrightarrow{N} \text{Loi LG} \left(\frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right); \frac{1}{\sqrt{n-3}} \right)$$

Cette transformation permet traiter le cas général, y compris lorsque le couple (X,Y) n'est pas une variable de Laplace-Gauss de dimension 2, dès que n est grand (n>30). Cependant le fait de ne pas rejeter l'hypothèse selon laquelle le coefficient de corrélation est nul, n'entraîne pas nécessairement l'indépendance des deux variables. Il n'y a là qu'une présomption d'indépendance. La nullité de ρ est une condition nécessaire mais pas suffisante pour l'indépendance. En d'autres mots l'absence de corrélation linéaire n'implique pas l'indépendance.

Le rapport de corrélation $\eta^2_{Y|X}$ de Y en X, le rapport de corrélation $\eta^2_{X|Y}$ de X en Y, et les rapports de corrélation empiriques

$$\eta^2_{Y|X} = \frac{V(m(Y|X))}{V(Y)} \quad \eta^2_{X|Y} = \frac{V(m(X|Y))}{V(X)}$$

$$(0 \leq \eta^2_{Y|X} \leq 1) \quad (0 \leq \eta^2_{X|Y} \leq 1)$$

Ce rapport est maximal et égal à 1 si la variable Y est **fonctionnellement**⁶ liée à la variable X. Il est à noter que ce coefficient est calculable que X soit une variable quantitative ou qualitative. Dans ce cas, nous ne pouvons calculer $\eta^2_{X|Y}$. Ces deux coefficients ne sont pas

symétriques. Le rapport de corrélation empirique s'obtient de la façon suivante :

$$E^2_{Y|X} = \frac{\frac{1}{n} \sum_{j=1}^k n_j (Y_j - Y_n)^2}{S_Y} \quad \text{avec} \quad S_{Y_n}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - Y_n)^2$$

Caractère significatif du rapport de corrélation :

Sous H_0 l'hypothèse de nullité de $\eta^2_{Y|X}$, on utilise la variable de décision :

$$D = \frac{\frac{E^2_{Y|X}}{k-1}}{1 - \frac{E^2_{Y|X}}{n-k}}$$

Sous H_0 l'hypothèse de nullité de $\eta^2_{Y|X}$, la variable de décision D suit la distribution de probabilité de la variable de Fisher-Snedecor de ddl (k-1;n-k) si les distributions conditionnelles de Y pour chaque valeurs ou modalité de X sont celles de la variable de

⁶ au sens mathématique du terme qui indique l'existence d'un lien univoque de X vers Y: à un résultat x de X ne correspond qu'un et un seul résultat y de Y. Si la réciproque est vraie, nous avons affaire à un lien biunivoque qui caractérise une fonction bijective.

Laplace-Gauss de même espérance μ et de même écart-type σ . Le nombre k correspond au nombre de valeurs ou de modalités de la variable X .