

Indépendance de deux variables et covariance

Jean-Claude Régnier

Professeur des Universités
Université Lumière Lyon2

Quelle information apporte la nullité de la covariance à l'égard de l'indépendance de deux variables quantitatives ? Quel intérêt peuvent avoir une représentation graphique et le rapport de corrélation $r^2_{Y/X}$?

Rappelons tout d'abord quelques définitions .

Soient X et Y deux variables statistiques de moyennes $m(X)$ et $m(Y)$, de variances $V(X)$ et $V(Y)$, ces deux variables sont dites indépendantes statistiquement si la réalisation de n'importe quel résultat pour X n'influence d'aucune façon celle d'un résultat quelconque pour Y. Cela peut se traduire par :

A étant un événement lié à X et B un événement lié à Y, la fréquence de (A et B) est égale au produit de la fréquence de (A) par la fréquence de (B) ou encore que la fréquence conditionnelle de (B sachant A) est égale à la fréquence de (A).

Par ailleurs la covariance est définie par la relation suivante :

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

Théorème :

Si les deux variables X et Y sont indépendantes statistiquement alors la covariance est nulle.

Cette propriété résulte immédiatement du fait que dans ce cas la moyenne du produit XY est égale au produit des moyennes respectives de X et de Y

Ainsi nous pouvons déduire logiquement le corollaire suivant :

Théorème :

Si les deux variables X et Y ont une covariance non-nulle alors elles ne sont pas indépendantes statistiquement (elles sont dépendantes statistiquement, cette dépendance ne doit pas être étendue abusivement à la dépendance causale).

Rappelons ensuite la prudence pour conclure.

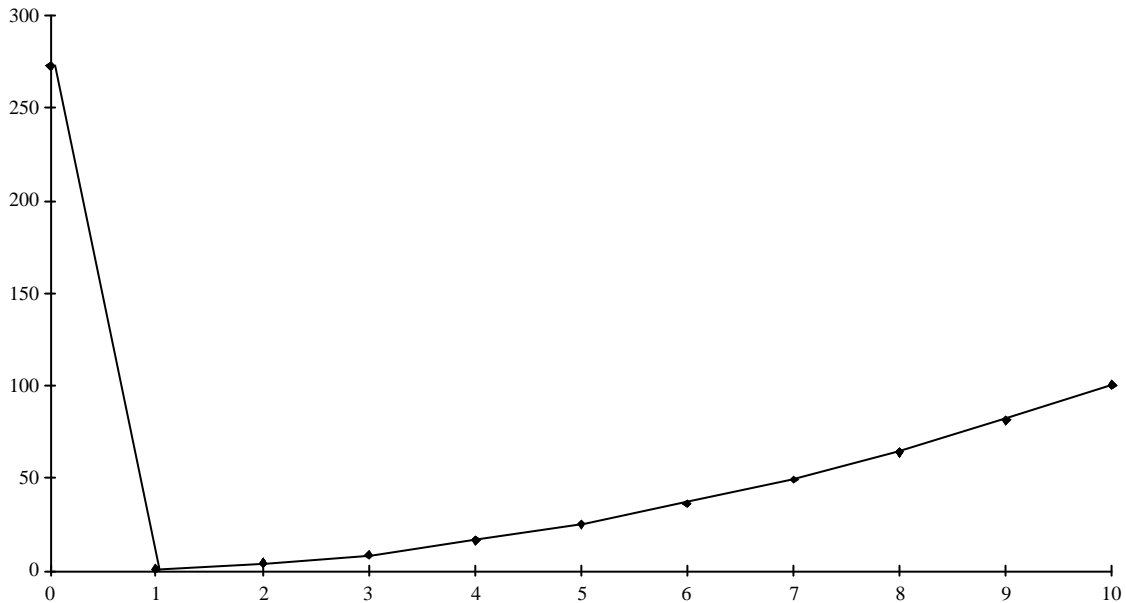
La covariance constitue alors un outil d'investigation en ce qui concerne le lien entre deux variables X et Y. Le problème est que si la non-nullité de la covariance implique l'existence d'un lien, sa nullité ne nous donne qu'une présomption d'indépendance. L'exemple que nous fournissons ci-dessous illustre la nécessaire prudence à maintenir dans les conclusions.

Le calcul montre que $\text{cov}(X, Y) = 278 - 5 \times 55,6 = 0$ alors que il existe une relation mathématique entre X et Y ainsi définie : $0 \rightarrow 272$ et $x \rightarrow y = x^2$

Ainsi en face d'un tel résultat, il serait particulièrement erroné de conclure à une indépendance des deux variables bien qu'elles soient non-corrélées et que

$$\rho(X, Y) = \frac{\text{cov}(X; Y)}{\sigma(X)\sigma(Y)} = 0$$

Indépendance & covariance



individus	X	Y	individus	X	Y
CE01	4	16	CE21	10	100
CE02	9	81	CE22	8	64
CE03	2	4	CE23	0	272
CE04	2	4	CE24	2	4
CE05	6	36	CE25	8	64
CE06	0	272	CE26	1	1
CE07	2	4	CE27	6	36
CE08	2	4	CE28	6	36
CE09	8	64	CE29	4	16
CE10	8	64	CE30	8	64
CE11	7	49	CE31	8	64
CE12	5	25	CE32	8	64
CE13	2	4	CE33	4	16
CE14	8	64	CE34	2	4
CE15	2	4	CE35	2	4
CE16	2	4	CE36	10	100
CE17	10	100	CE37	3	9
CE18	0	272	CE38	8	64
CE19	5	25	CE39	1	1
CE20	8	64	CE40	9	81

Le raisonnement suivant confirme le rejet de l'hypothèse d'indépendance entre les deux variables X et Y.

Y \ X	0	1	2	3	4	5	6	7	8	9	10	effectifs
1	0	2	0	0	0	0	0	0	0	0	0	2
4	0	0	10	0	0	0	0	0	0	0	0	10
9	0	0	0	1	0	0	0	0	0	0	0	1
16	0	0	0	0	3	0	0	0	0	0	0	3
25	0	0	0	0	0	2	0	0	0	0	0	2
36	0	0	0	0	0	0	3	0	0	0	0	3
49	0	0	0	0	0	0	0	1	0	0	0	1
64	0	0	0	0	0	0	0	0	10	0	0	10
81	0	0	0	0	0	0	0	0	0	2	0	2
100	0	0	0	0	0	0	0	0	0	0	3	3
272	3	0	0	0	0	0	0	0	0	0	0	3
effectifs	3	2	10	1	3	2	3	1	10	2	3	40

Ainsi $\text{Prop}\{X=2 \text{ et } Y=4\} = \frac{10}{40} = 0,025$ alors que $\text{Prop}\{X=2\} = \text{Prop}\{Y=4\} = \frac{10}{40}$.

Donc $\text{Prop}\{X=2 \text{ et } Y=4\} \neq \text{Prop}\{X=2\}\text{Prop}\{Y=4\}$.

Nous pouvons même remarquer que la relation d'indépendance n'est vérifiée par aucun des résultats du couple (X, Y). En effet si nous recourons à la fréquence conditionnelle, nous remarquons :

$\text{Prop}\{Y=y \text{ sachant } X=x\} = 1$ si $x=0$ et $y=272$ ou si $x=0$ et $y=x^2$.

$\text{Prop}\{Y=y \text{ sachant } X=x\} = 0$ si $x=0$ et $y \neq 272$ ou si $x \neq 0$ et $y \neq x^2$.

de là $\text{Prop}\{Y=y \text{ sachant } X=x\} \neq \text{Prop}\{Y=y\}$ pour toutes les valeurs de y de l'ensemble des résultats.

Pour analyser de plus près ce phénomène, nous pourrions recourir à un autre outil. Il s'agit du **rapport de corrélation de Y en X**, $\eta^2_{Y|X} = \frac{V(m(Y|X))}{V(Y)}$. Ce rapport est maximal et égal à 1 si la variable Y est **fonctionnellement**¹ liée à la variable X.

valeurs de X : k	0	1	2	3	4	5	6	7	8	9	10
moyennes conditionnelles de Y											
m(Y X=k)	272	1	4	9	16	25	36	49	64	81	100
fréquences	0,075	0,05	0,25	0,025	0,075	0,05	0,075	0,025	0,25	0,05	0,075

De ce tableau, il ressort que la moyenne de Y vaut 55,6 et que la variance de Y qui vaut $V(Y) = 4773,44$, est égale à la variance $V(m(Y|X))$ des moyennes conditionnelles de Y. Si

¹ 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.

nous calculons maintenant le rapport de corrélation de Y en X, $\eta = \sqrt{\frac{V(m(Y|X))}{V(Y)}}$, nous obtenons la valeur 1 qui révèle l'existence de la liaison fonctionnelle de Y avec X lisible sur la représentation graphique.

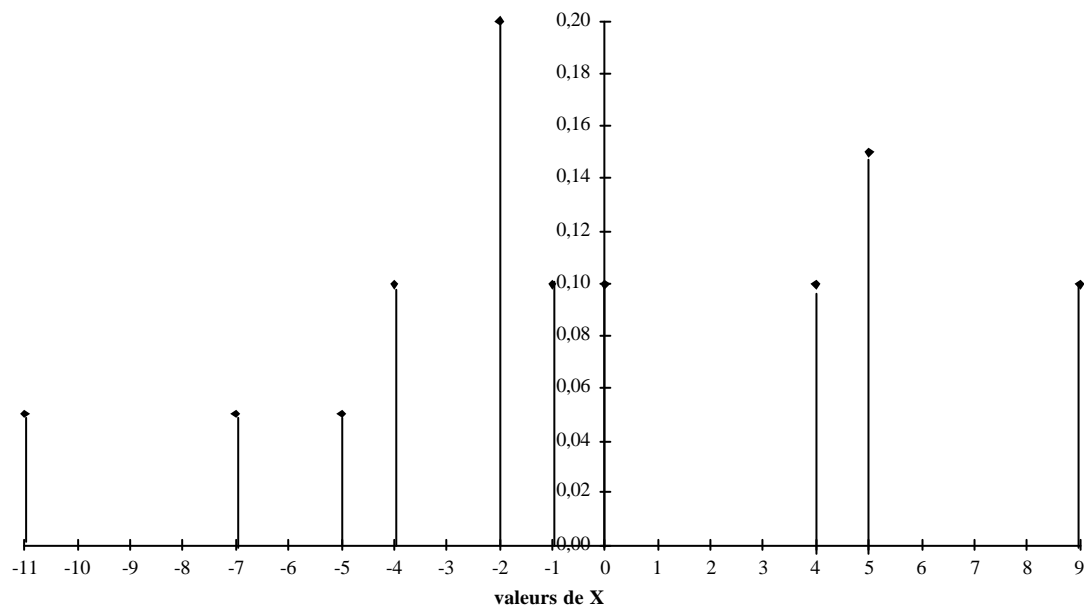
Nous pourrions obtenir un résultat plus général en considérant **X comme une variable centrée et symétrique** car alors $m(X) = 0$ et le moment centré d'ordre 3,

$$m[(X-0)^3] = m[X^3] = 0. \text{ De là si nous considérons la variable } Y = X^2 \text{ nous obtenons alors } \\ \text{cov}(X, Y) = m[XY] - m[X] m[Y] = m[X^3] - m[X] m[X^2] = 0.$$

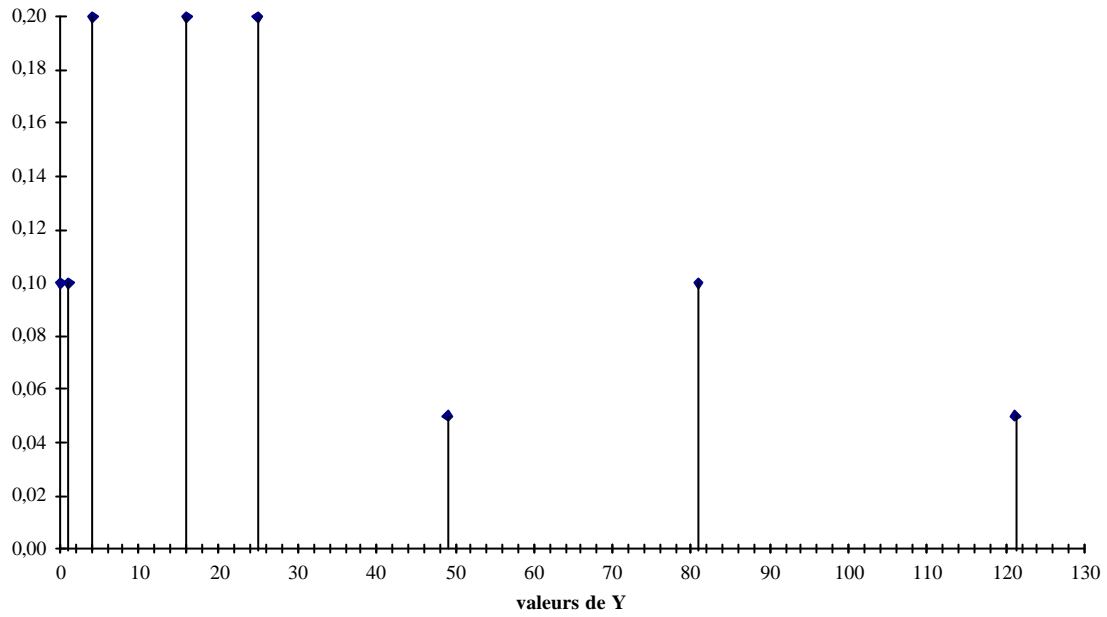
Cependant la condition de symétrie n'est pas nécessaire comme l'illustre l'exemple que nous donnons ci-après.

X \ Y	-11	-7	-5	-4	-2	-1	0	4	5	9	effectifs
0	0	0	0	0	0	0	4	0	0	0	4
1	0	0	0	0	0	4	0	0	0	0	4
4	0	0	0	0	8	0	0	0	0	0	8
16	0	0	0	4	0	0	0	4	0	0	8
25	0	0	2	0	0	0	0	0	6	0	8
49	0	2	0	0	0	0	0	0	0	0	2
81	0	0	0	0	0	0	0	0	0	4	4
121	2	0	0	0	0	0	0	0	0	0	2
effectifs	2	2	2	4	8	4	4	4	6	4	40

distribution de fréquence de X



distribution de fréquence de Y

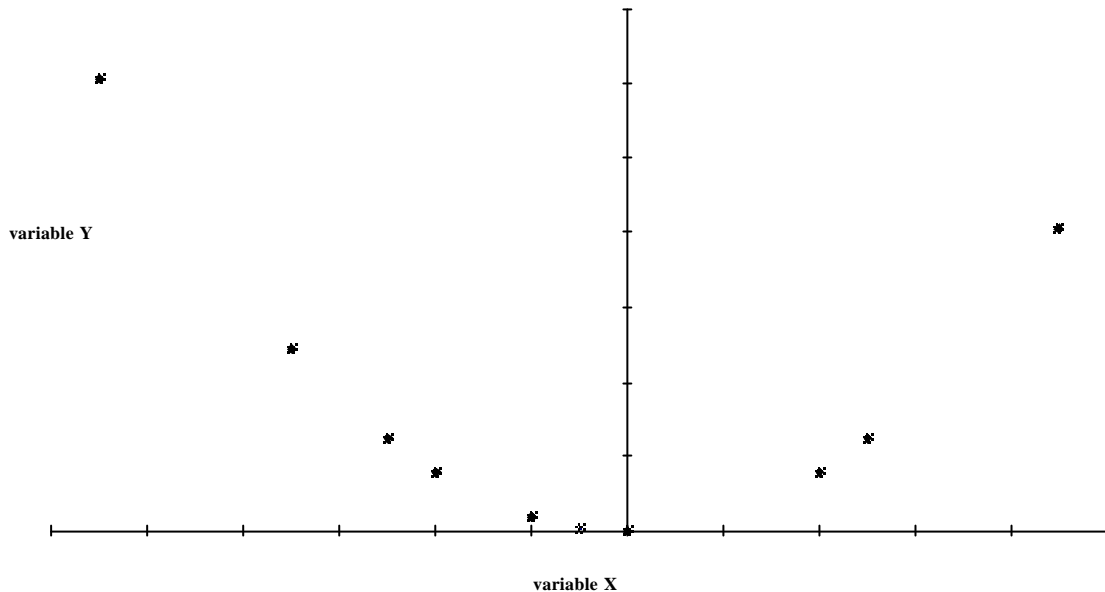


Ci-après nous rappelons le tableau des séries statistiques de X et de Y.

individus	X	Y	individus	X	Y
I001	-11	121	I021	-1	1
I002	-11	121	I022	-1	1
I003	-7	49	I023	0	0
I004	-7	49	I024	0	0
I005	-5	25	I025	0	0
I006	-5	25	I026	0	0
I007	-4	16	I027	4	16
I008	-4	16	I028	4	16
I009	-4	16	I029	4	16
I010	-4	16	I030	4	16
I011	-2	4	I031	5	25
I012	-2	4	I032	5	25
I013	-2	4	I033	5	25
I014	-2	4	I034	5	25
I015	-2	4	I035	5	25
I016	-2	4	I036	5	25
I017	-2	4	I037	9	81
I018	-2	4	I038	9	81
I019	-1	1	I039	9	81
I020	-1	1	I040	9	81

Nuage des points de coordonnées (x, y). Certains sont superposés

indépendance & covariance



Ici $m(X) = 0$ mais aussi $m[X^3] = 0$. De là si nous considérons la variable $Y = X^2$ nous obtenons encore $\text{cov}(X, Y) = m[XY] - m[X]m[Y] = m[X^3] - m[X]m[X^2] = 0$.

Ainsi $\text{Prop}\{X=5 \text{ et } Y=25\} = \frac{6}{40}$ alors que $\text{Prop}\{X=5\} = \frac{6}{40}$ et $\text{Prop}\{Y=25\} = \frac{8}{40}$.

Donc $\text{Prop}\{X=5 \text{ et } Y=25\} < \text{Prop}\{X=5\}\text{Prop}\{Y=25\}$. Pour que ces deux événements soient indépendants, il aurait fallu obtenir $\text{Prop}\{X=5 \text{ et } Y=25\} = 0,03$. Quoi qu'il en soit ceci conduit à rejeter l'hypothèse d'indépendance de X et de Y pourtant suggérée par la nullité de la covariance.

Tout cela confirme la nécessité de grande prudence quant à l'information issue de la nullité de la covariance ou du coefficient de corrélation linéaire qui en découle pour porter un jugement à l'égard de l'indépendance de deux variables quantitatives. Qui plus est, quand la nullité de la covariance est elle-même le résultat d'un test statistique conduisant à ne pas rejeter l'hypothèse H_0 au niveau α , c'est à dire à conserver H_0 avec un risque de second espèce de niveau β , nous pouvons imaginer combien la conservation de l'hypothèse d'indépendance est risquée. Toutefois en utilisant une représentation graphique pour repérer les points de coordonnées $(x;y)$ et le rapport de corrélation $\eta^2_{Y|X}$, nous pouvons confronter la vraisemblance de la présomption d'indépendance de X et de Y à une forme remarquable du nuage de points. A cela s'ajoute d'autres éclairages que le contexte dans lequel le problème de recherche de liaison entre X et Y a été posé de manière pertinente, doit pouvoir apporter.

valeurs de X : k	-11	-7	-5	-4	-2	-1	0	4	5	9
moyennes conditionnelles de Y										
m(Y X=k)	121	49	25	16	4	1	0	16	25	81
fréquences	0,05	0,05	0,05	0,1	0,2	0,1	0,1	0,1	0,15	0,1
valeurs de Y : k	0	1	4	16	25	49	81	121		
moyennes conditionnelles de X										
m(X Y=k)	0	-1	-2	0	2,5	-7	9	-11		
fréquences	0,1	0,1	0,2	0,2	0,2	0,05	0,1	0,05		

Les calculs des divers paramètres utiles nous fournissent les résultats suivants :

$$m(X) = m[m(X|Y)] = 0$$

$$m(Y) = m[m(Y|X)] = 25,7$$

$$V(X) = 25,7$$

$$V(Y) = 1027,21$$

$$V[m(X|Y)] = 18,75$$

$$V[m(Y|X)] = 1027,21$$

D'où d'une part la covariance $cov(X, Y) = 0$ et par conséquent le coefficient de corrélation (linéaire) $\rho = 0$, d'autre part les deux rapports de corrélation prennent respectivement les valeurs suivantes :

rapport de corrélation de Y en X, $\eta^2_{Y|X}$

rapport de corrélation de X en Y, $\eta^2_{X|Y}$

$$= \frac{V(m(Y|X))}{V(Y)} = \frac{1027,21}{1027,21} = 1$$

$$= \frac{V(m(X|Y))}{V(X)} = \frac{18,75}{25,7} = 0,729$$

Ainsi la nullité du coefficient de corrélation (linéaire) traduit l'absence de liaison linéaire et non l'indépendance des deux variables X et Y. Par ailleurs la valeur du rapport de corrélation de Y en X, $\eta^2_{Y|X} = 1$, confirme l'existence d'une liaison fonctionnelle, ici $Y = X^2$, alors que celle du rapport de corrélation de X en Y, $\eta^2_{X|Y} = 0,729$, rend compte d'une liaison non-fonctionnelle.

Cet exemple confirme toute l'attention que nous devons porter à la formulation des conclusions relatives aux recherches de liaisons, quand le raisonnement prend appui sur ces outils. L'étude théorique de ces outils mathématiques paraît être un bon moyen pour mieux maîtriser ces formulations.