Covariance de deux variables aléatoires

Dans une urne contenant n jetons indiscernables au toucher (n >1), on tire successivement deux jetons, et on considère les variables aléatoires X et Y qui prennent respectivement les valeurs successives des numéros obtenus.

Ainsi, si la suite de numéros est formée du couple (1 ; 2), X vaut 1 et Y vaut 2.

Le but est d'évaluer la corrélation des variables X et Y en commençant par évaluer leur covariance.

Rappels :

La **covariance** de deux variables aléatoires réelles X et Y se définit comme étant :

$$\mathrm{cov}(X\,;\,Y) = E(XY) - E(X)\,E(Y)$$

Elle permet de calculer le **coefficient de corrélation linéaire** des deux variables, défini par :

$$r(X\,;Y) = \frac{\mathrm{cov}(X\,Y)}{\sigma(X)\,\sigma(Y)}$$

où $\sigma(X)$ et $\sigma(Y)$ désignent les écarts types des variables.

1ère étape : Loi de probabilité conjointe :

Pour résoudre l'exercice, il faut commencer par déterminer la loi de probabilité conjointe des variables X et Y, c'est-à-dire, pour tous les couples (i, j) d'entiers de $[\![1; n]\!]$:

$$P\big((X = i) \cap (Y = j)\big)$$

Or, le tirage se faisant sans remise :

$$\text{si } i \neq j : \quad P\big((X = i) \cap (Y = j)\big) = \frac{1}{n\,(n-1)}$$

$$\text{si } i = j : \quad P\big((X = i) \cap (Y = j)\big) = 0$$

2ème étape : Cacul de E(X Y) :

Pour ce faire, nous allons dresser le tableau des résultats possibles de la variable $Z = X\,Y$, selon les occurrences de X et de Y.

X\Y	1	2	3	j	…	…	n
1		2	3	j	…	…	n
2	2		…	2 j	…	…	2 n
3	3	6		⋮	…	…	⋮
⋮	⋮	⋮	…		…	…	⋮
i	i	2 i	…	i j		…	i n
⋮	⋮	⋮	… …	⋮	…		⋮
n	n	2 n	…	n j	…	…	

Notez alors que ce tableau est symétrique par rapport à sa diagonale descendante. Les cellules de couleur rouge correspondent à des occurrences impossibles pour le couple (i ; j)

Nous allons alors calculer l'espérance en utilisant la symétrie du tableau et en faisant une sommation par colonnes sur la partie en bleue du tableau :

X\Y	1	2	3	j	…	…	n
1		2	3	j	…	…	n
2	2		…	2 j	…	…	2 n
3	3	6		⋮	…	…	⋮
⋮	⋮	⋮	…		…	…	⋮
i	i	2 i	…	i j		…	i n
⋮	⋮	⋮	… …	⋮	…		⋮
n	n	2 n	…	n j	…	…	

Les colonnes étant indexées par j variant de 1 à (n-1), nous avons :

$$E(X\,Y) = 2\sum_{j=1}^{n-1}\sum_{i=j+1}^{n}\left(i\,j\,P\big((X=i)\ \cap (Y=j)\big)\right)$$

$$= 2\sum_{j=1}^{n-1}\sum_{i=j+1}^{n}\left(i\,j\,\frac{1}{n\,(n-1)}\right)$$

$$= \frac{2}{n\,(n-1)}\sum_{j=1}^{n-1}\left(j\sum_{i=j+1}^{n} i\right)$$

Or la somme sur i est une somme de termes consécutifs d'une suite arithmétique de raison 1 dont nous rappelons la formule :

$$\text{somme} = \frac{(\text{premier terme} + \text{dernier terme}) \times \text{nombre de terme}}{2}$$

On en déduit, le nombre de termes étant $(n-j)$:

$$E(X\,Y) = \frac{2}{n\,(n-1)}\sum_{j=1}^{n-1}\left(j\,\frac{(j+1+n)\,(n-j)}{2}\right)$$

$$= \frac{1}{n\,(n-1)}\sum_{j=1}^{n-1}\big(j\,(j+1+n)\,(n-j)\big)$$

$$= \frac{1}{n\,(n-1)}\sum_{j=1}^{n-1}(j\,(n\,j-j^2+n-j+n^2-n\,j)\,)$$

$$= \frac{1}{n\,(n-1)} \sum_{j=1}^{n-1} (-j^3 - j^2 + (n^2 + n)\,j)$$

$$= \frac{-1}{n\,(n-1)} \sum_{j=1}^{n-1} j^3 - \frac{1}{n\,(n-1)} \sum_{j=1}^{n-1} j^2 + \frac{n+1}{n-1} \sum_{j=1}^{n-1} j$$

Rappelons alors les sommes remarquables :

$$\sum_{j=1}^{N} j = \frac{N\,(N+1)}{2}$$

$$\sum_{j=1}^{N} j^2 = \frac{N\,(N+1)\,(2\,N+1)}{6}$$

$$\sum_{j=1}^{N} j^3 = \left(\frac{N\,(N+1)}{2}\right)^2$$

On en déduit, en remplaçant N par $(n-1)$:

$$E(X\,Y) = \frac{-1}{n\,(n-1)} \frac{\left(n\,(n-1)\right)^2}{4} - \frac{1}{n\,(n-1)} \frac{n\,(n-1)\,(2\,n-1)}{6} + \frac{n+1}{n-1} \times \frac{n\,(n-1)}{2}$$

$$= \frac{-n\,(n-1)}{4} - \frac{(2\,n-1)}{6} + \frac{n\,(n+1)}{2}$$

$$= \frac{-3\,n\,(n-1) - 2\;(2\,n-1) + 6\,n\,(n+1)}{12}$$

Finalement :

$$\boxed{E(X\,Y) = \frac{3\,n^2 + 5n + 2}{12}}$$

3ème étape : Cacul de E(X) et de E(Y) :

Faisons le tableau de la loi conjointe des deux variables X et Y en notant :

$$p = \frac{1}{n\,(n-1)}$$

X\Y	1	2	3	j	…	…	n
1		p	p	p	…	…	p
2	p		…	p	…	…	p
3	p	p		⋮	…	…	⋮
⋮	⋮	⋮	…		…	…	⋮
i	p	p	…	p		…	p
⋮	⋮	⋮	… …	⋮	…		⋮
n	p	p	…	p	…	…	

On en déduit facilement les lois marginales de X et Y :

En sommant les éléments de la ligne i du tableau, nous avons en effet :

$$P(X = i) = (n-1)\,p = \frac{1}{n}$$

Et en sommant les éléments de la colonne j :

$$P(Y = j) = (n-1)\,p = \frac{1}{n}$$

Les espérances s'en déduisent :

$$E(X) = \sum_{i=1}^{n} i\,P(X = i) = \frac{1}{n}\sum_{i=1}^{n} i = \frac{1}{n} \times \frac{n\,(n+1)}{2}$$

Soit :

$$E(X) = \frac{(n+1)}{2}$$

De même :

$$E(Y) = \sum_{j=1}^{n} j\, P(X = j) = \frac{(n+1)}{2}$$

4ème étape : Cacul de Cov(X ; Y):

$$Cov(X\,;Y) = E(X\,Y) - E(X)\,E(Y)$$

$$Cov(X\,;Y) = \frac{3\,n^2 + 5n + 2}{12} - \frac{(n+1)^2}{4}$$

$$= \frac{3\,n^2 + 5n + 2 - 3\,(n^2 + 2n + 1)}{12}$$

finalement :

$$Cov(X\,;Y) = \frac{-\,n - 1}{12}$$

5ème étape : Cacul de V(X) et de V(Y) :

Commençons par l'espérance des carrés de X et Y :

$$E(X^2) = \sum_{i=1}^{n} i^2\, P(X = i) = \frac{1}{n}\sum_{i=1}^{n} i^2 = \frac{1}{n} \times \frac{n\,(n+1)(2\,n+1)}{6}$$

$$E(X^2) = \frac{(n+1)(2\,n+1)}{6} = E(Y^2)$$

Les variances s'en déduisent :

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

$$= \frac{(n+1)(2\,n+1)}{6} - \frac{(n+1)^2}{4}$$

$$= \frac{2\,(2\,n^2 + 3\,n + 1) - 3(n^2 + 2\,n + 1)}{12}$$

Finalement :

$$V(X) = \frac{n^2 - 1}{12} = V(Y)$$

Les écarts types s'en déduisent :

$$\sigma(X) = \sqrt{\frac{n^2 - 1}{12}} = \sigma(Y)$$

<u>6ème étape : Cacul du coefficient de corrélation linéaire</u> :

$$r(X\,Y) = \frac{cov(X\,Y)}{\sigma(X)\,\sigma(Y)}$$

$$= \frac{\frac{-(n+1)}{12}}{\sqrt{\frac{n^2 - 1}{12}}\sqrt{\frac{n^2 - 1}{12}}}$$

$$= \frac{-(n+1)}{n^2 - 1}$$

$$= \frac{-(n+1)}{(n+1)(n-1)}$$

Finalement :

$$r(X;\ Y) = \frac{-1}{(n-1)}$$

Nous voyons donc que les deux variables sont faiblement corrélées et négativement :

Pour $n = 2$, r vaut – 50% et pour $n = 11$, -10%. Les deux variables sont d'autant moins corrélées que n est grand.

Le coefficient de corrélation tend vers 0 quand n tend vers l'infini. Pour des n suffisamment grands (supérieur à 11 environ) la corrélation est inférieure à 10 % en valeur absolue. Cela signifie que si on représentait un grand nombre de réalisations (i ; j) par un nuage de points (i en abscisse et j en ordonnée), le nuage s'envelopperait dans une sorte de disque. Alors que pour corrélation de -50%, l'enveloppe serait une sorte d'ellipse dont le grand axe serait incliné vers le bas.

Rappelons qu'un coefficient de corrélation de 100% (1) donnerait un nuage de points alignés sur une droite de coefficient directeur strictement positif et un coefficient de – 100%, un nuage de points alignés sur une droite de coefficient directeur strictement négatif, ce qui traduirait un lien affine entre les variables X et Y ($Y = a X + b$) mais ce n'était pas le cas traité dans notre exercice.