DM 6 Espérance et corrélation

Exercice 1 – (Minimum et maximum de deux variables géométriques) :

Soit $p \in]0\,;1[$ et $q = 1-p$. On pose $r = 1-q^2$.

Soient $\mathbb{X}_1$ et $\mathbb{X}_2$ deux variables aléatoires indépendantes, de même loi géométrique $\mathcal{G}(p)$. On définit

$$\mathbb{U} = \min(\mathbb{X}_1, \mathbb{X}_2) \quad \text{et} \quad \mathbb{V} = \max(\mathbb{X}_1, \mathbb{X}_2).$$

Partie I – Loi de $\mathbb{U}$

1. Rappeler l'espérance et la variance de $\mathbb{X}_1$. Montrer que, pour tout $n \in \mathbb{N}$, $\mathbf{P}([\mathbb{X}_1 > n]) = q^n$.

   On sait que $\mathbf{E}(\mathbb{X}_1) = \dfrac{1}{p}$ et $\mathbf{V}(\mathbb{X}_1) = \dfrac{q}{p^2}$.
   Soit $n \in \mathbb{N}$. Comme $\mathbb{X}_1(\Omega) = \mathbb{N}^*$ et $\mathbf{P}([\mathbb{X}_1 = k]) = pq^{k-1}$ pour tout $k \geqslant 1$,

   $$\mathbf{P}([\mathbb{X}_1 > n]) = \mathbf{P}\left(\bigcup_{k=n+1}^{+\infty} [\mathbb{X}_1 = k]\right) = \sum_{k=n+1}^{+\infty} \mathbf{P}([\mathbb{X}_1 = k]) = \sum_{k=n+1}^{+\infty} pq^{k-1} = p\,\frac{q^n}{1-q} = q^n.$$

2. Soit $n \in \mathbb{N}$.
   Exprimer l'événement $[\mathbb{U} > n]$ à l'aide de $[\mathbb{X}_1 > n]$ et $[\mathbb{X}_2 > n]$.
   En déduire $\mathbf{P}([\mathbb{U} > n])$, puis que $\mathbb{U} \rightsquigarrow \mathcal{G}(r)$. Donner $\mathbf{E}(\mathbb{U})$ et $\mathbf{V}(\mathbb{U})$.

   On a $[\mathbb{U} > n] = [\mathbb{X}_1 > n] \cap [\mathbb{X}_2 > n]$.
   Par indépendance de $\mathbb{X}_1$ et $\mathbb{X}_2$ et d'après la question précédente,

   $$\mathbf{P}([\mathbb{U} > n]) = \mathbf{P}([\mathbb{X}_1 > n])\,\mathbf{P}([\mathbb{X}_2 > n]) = q^{2n}.$$

   Alors $\mathbb{U}(\Omega) = \mathbb{N}^*$ et, pour tout $n \geqslant 1$,

   $$\mathbf{P}([\mathbb{U} = n]) = \mathbf{P}([\mathbb{U} > n-1]) - \mathbf{P}([\mathbb{U} > n]) = q^{2(n-1)}\,(1-q^2) = r\,(1-r)^{n-1}.$$

   Comme $r = 1-q^2 \in]0\,;1[$, on reconnaît $\mathbb{U} \rightsquigarrow \mathcal{G}(r)$.
   Par conséquent,

   $$\mathbf{E}(\mathbb{U}) = \frac{1}{r} \quad \text{et} \quad \mathbf{V}(\mathbb{U}) = \frac{1-r}{r^2} = \frac{q^2}{r^2}.$$

Partie II – Espérance et covariance

3. Vérifier que $\mathbb{U} + \mathbb{V} = \mathbb{X}_1 + \mathbb{X}_2$ et $\mathbb{U}\mathbb{V} = \mathbb{X}_1\mathbb{X}_2$. En déduire que $\mathbb{V}$ admet une espérance et que

   $$\mathbf{E}(\mathbb{V}) = \frac{2}{p} - \frac{1}{r}.$$

Pour tout $\omega \in \Omega$, $\{\min(\mathbb{X}_1(\omega), \mathbb{X}_2(\omega)), \max(\mathbb{X}_1(\omega), \mathbb{X}_2(\omega))\} = \{\mathbb{X}_1(\omega), \mathbb{X}_2(\omega)\}$: la somme et le produit de ces deux nombres sont donc ceux de $\mathbb{X}_1(\omega)$ et $\mathbb{X}_2(\omega)$. D'où

$$\mathbb{U} + \mathbb{V} = \mathbb{X}_1 + \mathbb{X}_2 \quad \text{et} \quad \mathbb{U}\mathbb{V} = \mathbb{X}_1\mathbb{X}_2.$$

Ainsi $\mathbb{V} = \mathbb{X}_1 + \mathbb{X}_2 - \mathbb{U}$ est une combinaison linéaire de variables aléatoires admettant une espérance, donc $\mathbb{V}$ admet une espérance et, par linéarité,

$$\mathbf{E}(\mathbb{V}) = \mathbf{E}(\mathbb{X}_1) + \mathbf{E}(\mathbb{X}_2) - \mathbf{E}(\mathbb{U}) = \frac{2}{p} - \frac{1}{r}.$$

4. Montrer que $\mathbb{U}$ et $\mathbb{V}$ admettent un moment d'ordre 2. Montrer ensuite que

$$\mathbf{Cov}(\mathbb{U}, \mathbb{V}) = \frac{q^2}{r^2}.$$

— Les variables $\mathbb{X}_1$ et $\mathbb{X}_2$ admettent une variance, donc un moment d'ordre 2, ainsi que leur somme. Comme $0 \leqslant \mathbb{U} \leqslant \mathbb{V} \leqslant \mathbb{X}_1 + \mathbb{X}_2$, on a $0 \leqslant \mathbb{U}^2 \leqslant \mathbb{V}^2 \leqslant (\mathbb{X}_1 + \mathbb{X}_2)^2$.
Par domination, $\mathbb{U}^2$ et $\mathbb{V}^2$ admettent une espérance : $\mathbb{U}$ et $\mathbb{V}$ admettent un moment d'ordre 2.
La covariance de $(\mathbb{U}, \mathbb{V})$ existe donc.

— D'après la question précédente et l'indépendance de $\mathbb{X}_1$ et $\mathbb{X}_2$,

$$\mathbf{E}(\mathbb{U}\mathbb{V}) = \mathbf{E}(\mathbb{X}_1\mathbb{X}_2) = \mathbf{E}(\mathbb{X}_1)\,\mathbf{E}(\mathbb{X}_2) = \frac{1}{p^2}.$$

Par la formule de Koenig-Huygens,

$$\begin{aligned}\mathbf{Cov}(\mathbb{U}, \mathbb{V}) &= \mathbf{E}(\mathbb{U}\mathbb{V}) - \mathbf{E}(\mathbb{U})\mathbf{E}(\mathbb{V}) \\ &= \frac{1}{p^2} - \frac{1}{r}\left(\frac{2}{p} - \frac{1}{r}\right) = \frac{1}{p^2} - 2\frac{1}{r}\frac{1}{p} + \frac{1}{r^2} \\ &= \left(\frac{1}{p} - \frac{1}{r}\right)^2.\end{aligned}$$

Or $r - p = p(2-p) - p = p(1-p) = pq$, d'où $\dfrac{1}{p} - \dfrac{1}{r} = \dfrac{r-p}{pr} = \dfrac{q}{r}$.
Finalement,

$$\mathbf{Cov}(\mathbb{U}, \mathbb{V}) = \frac{q^2}{r^2}.$$

5. Établir que $\mathbf{Cov}(\mathbb{U}, \mathbb{V}) = \mathbf{V}(\mathbb{U})$. En déduire que $\mathbb{U}$ et $\mathbb{V} - \mathbb{U}$ sont non corrélées.

D'après les questions 2 et 4, $\mathbf{Cov}(\mathbb{U}, \mathbb{V}) = \dfrac{q^2}{r^2} = \mathbf{V}(\mathbb{U})$.
Par linéarité à droite de la covariance,

$$\mathbf{Cov}(\mathbb{U}, \mathbb{V} - \mathbb{U}) = \mathbf{Cov}(\mathbb{U}, \mathbb{V}) - \mathbf{Cov}(\mathbb{U}, \mathbb{U}) = \mathbf{Cov}(\mathbb{U}, \mathbb{V}) - \mathbf{V}(\mathbb{U}) = 0.$$

Les variables $\mathbb{U}$ et $\mathbb{V} - \mathbb{U}$ sont donc non corrélées.

6. Les variables aléatoires $\mathbb{U}$ et $\mathbb{V}$ sont-elles indépendantes ?

Non. Puisque $q \in]0\,;1[$, $\mathbf{Cov}(\mathbb{U}, \mathbb{V}) = \dfrac{q^2}{r^2} \neq 0$.
Comme deux variables aléatoires indépendantes admettant un moment d'ordre 2 ont une covariance nulle, alors par la contraposée, $\mathbb{U}$ et $\mathbb{V}$ ne sont pas indépendantes.

Partie III – Variance de $\mathbb{V}$ et corrélation

7. Calculer $\mathbf{V}(\mathbb{U}+\mathbb{V})$ de deux façons différentes. En déduire

$$\mathbf{V}(\mathbb{V}) = \frac{2q}{p^2} - \frac{3q^2}{r^2}.$$

— Comme $\mathbb{U}+\mathbb{V} = \mathbb{X}_1 + \mathbb{X}_2$ avec $\mathbb{X}_1$ et $\mathbb{X}_2$ indépendantes admettant une variance,

$$\mathbf{V}(\mathbb{U}+\mathbb{V}) = \mathbf{V}(\mathbb{X}_1) + \mathbf{V}(\mathbb{X}_2) = \frac{2q}{p^2}.$$

— Par ailleurs, $\mathbf{V}(\mathbb{U}+\mathbb{V}) = \mathbf{V}(\mathbb{U}) + \mathbf{V}(\mathbb{V}) + 2\,\mathbf{Cov}(\mathbb{U},\mathbb{V}) = \mathbf{V}(\mathbb{V}) + \dfrac{3q^2}{r^2}$, d'après la question 5.

En comparant, $\mathbf{V}(\mathbb{V}) = \mathbf{V}(\mathbb{U}+\mathbb{V}) - \dfrac{3q^2}{r^2} = \dfrac{2q}{p^2} - \dfrac{3q^2}{r^2}$.

8. Montrer que $r = p(1+q)$, puis que le coefficient de corrélation linéaire de $(\mathbb{U},\mathbb{V})$ vaut

$$\boldsymbol{\rho}(\mathbb{U},\mathbb{V}) = \sqrt{\frac{q}{2q^2+q+2}}.$$

— On a $r = 1-q^2 = (1-q)(1+q) = p(1+q)$.
Ainsi,

$$\mathbf{V}(\mathbb{V}) = \frac{2q}{p^2} - \frac{3q^2}{p^2(1+q)^2} = \frac{q\,(2(1+q)^2 - 3q)}{p^2(1+q)^2} = \frac{q\,(2q^2+q+2)}{p^2(1+q)^2}.$$

— Les variances de $\mathbb{U}$ et $\mathbb{V}$ sont non nulles, donc $\boldsymbol{\rho}(\mathbb{U},\mathbb{V})$ est bien défini.
Comme $\mathbf{Cov}(\mathbb{U},\mathbb{V}) = \mathbf{V}(\mathbb{U})$,

$$\boldsymbol{\rho}(\mathbb{U},\mathbb{V}) = \frac{\mathbf{Cov}(\mathbb{U},\mathbb{V})}{\sqrt{\mathbf{V}(\mathbb{U})\,\mathbf{V}(\mathbb{V})}} = \frac{\mathbf{V}(\mathbb{U})}{\sqrt{\mathbf{V}(\mathbb{U})\,\mathbf{V}(\mathbb{V})}} = \sqrt{\frac{\mathbf{V}(\mathbb{U})}{\mathbf{V}(\mathbb{V})}}.$$

Or $\mathbf{V}(\mathbb{U}) = \dfrac{q^2}{r^2} = \dfrac{q^2}{p^2(1+q)^2}$, d'où

$$\boldsymbol{\rho}(\mathbb{U},\mathbb{V}) = \sqrt{\frac{\cancel{\frac{q^2}{p^2(1+q)^2}}}{\cancel{\frac{q\,(2q^2+q+2)}{p^2(1+q)^2}}}} = \sqrt{\frac{q}{2q^2+q+2}}.$$

9. En étudiant la fonction $f : q \longmapsto \dfrac{q}{2q^2+q+2}$ pour $q \in \,]0\,;1[$, vérifier que $0 < \boldsymbol{\rho}(\mathbb{U},\mathbb{V}) < \dfrac{1}{\sqrt{5}}$ et commenter.

Son discriminant étant strictement négatif, le trinôme $2q^2+q+2$ est de signe constant strictement positif et $f(q) > 0$, donc $\boldsymbol{\rho}(\mathbb{U},\mathbb{V}) > 0$ (ce qui est cohérent avec le signe de la covariance).
De plus, quotient de fonctions dérivables dont le dénominateur ne s'annule pas sur $]0\,;1[$, f l'est également et on a : f est dérivable et

$$\forall\, q \in\,]0\,;1[\,,\ f'(q) = \frac{(2q^2+q+2) - q(4q+1)}{(2q^2+q+2)^2} = \frac{2(1-q^2)}{(2q^2+q+2)^2} > 0.$$

Ainsi f est strictement croissante sur $]0\,;1[$.

En particulier, $\forall\, q \in]0\,;1[,\ f(q) < f(1) = \dfrac{1}{5}$.

D'où $0 < \boldsymbol{\rho}(\mathbb{U},\mathbb{V}) < \dfrac{1}{\sqrt{5}} \approx 0{,}45$.

Commentaire : $\mathbb{U}$ et $\mathbb{V}$ sont positivement corrélées (elles évoluent en moyenne dans le même sens) mais la dépendance linéaire entre elles reste faible, quelle que soit la valeur de p.

Partie IV – Meilleure approximation affine et simulation

10. Pour tout $t \in \mathbb{R}$, exprimer $\mathbf{V}(\mathbb{V} - t\,\mathbb{U})$ en fonction de t, q et r.
Pour quelle valeur de t cette quantité est-elle minimale ? Donner alors ce minimum.

Soit $t \in \mathbb{R}$. Les variables $\mathbb{U}$ et $\mathbb{V}$ admettent une variance, donc $\mathbb{V} - t\,\mathbb{U}$ aussi et, par linéarité à gauche de **Cov**, on a :

$$\begin{aligned}\mathbf{V}(\mathbb{V} - t\,\mathbb{U}) &= \mathbf{V}(\mathbb{V}) - 2\,\mathbf{Cov}(t\mathbb{U},\mathbb{V}) + \mathbf{V}(t\mathbb{U})\\ &= \mathbf{V}(\mathbb{V}) - 2t\,\mathbf{Cov}(\mathbb{U},\mathbb{V}) + t^2\mathbf{V}(\mathbb{U})\\ &= \mathbf{V}(\mathbb{V}) + \frac{q^2}{r^2}\left(t^2 - 2t\right).\end{aligned}$$

Or, la fonction $t \longmapsto t^2 - 2t = (t-1)^2 - 1$ est une fonction du second degré admettant un minimum pour $t = 1$ égal à -1.
La quantité $\mathbf{V}(\mathbb{V} - t\,\mathbb{U})$ est donc minimale pour $t = 1$ de minimum

$$\mathbf{V}(\mathbb{V} - \mathbb{U}) = \mathbf{V}(\mathbb{V}) - \frac{q^2}{r^2} = \frac{2q}{p^2} - \frac{4q^2}{r^2}.$$

Remarque : C'est la variance de $\mathbb{V} - \mathbb{U} = |\mathbb{X}_1 - \mathbb{X}_2|$, qui est ainsi la « meilleure » correction de $\mathbb{V}$ par un multiple de $\mathbb{U}$.

11. Pour $p = \dfrac{1}{2}$, calculer $\mathbf{E}(\mathbb{U})$, $\mathbf{E}(\mathbb{V})$ et $\mathbf{Cov}(\mathbb{U},\mathbb{V})$.
Écrire ensuite un script `Python` simulant un grand nombre de réalisations de $(\mathbb{U},\mathbb{V})$ et affichant les moyennes empiriques de $\mathbb{U}$ et $\mathbb{V}$ ainsi que leur covariance empirique, afin de comparer avec les valeurs théoriques.

Pour $p = q = \dfrac{1}{2}$, $r = \dfrac{3}{4}$ et

$$\mathbf{E}(\mathbb{U}) = \frac{4}{3}, \qquad \mathbf{E}(\mathbb{V}) = 4 - \frac{4}{3} = \frac{8}{3}, \qquad \mathbf{Cov}(\mathbb{U},\mathbb{V}) = \frac{\frac{1}{4}}{\frac{9}{16}} = \frac{4}{9}.$$

Un script possible (la fonction `rd.geometric(p, N)` simule N variables de loi $\mathcal{G}(p)$) :

Éditeur
```
import numpy as np
import numpy.random as rd
p = 0.5
N = 10**5
X1 = rd.geometric(p, N)
```

Éditeur
```
X2 = rd.geometric(p, N)
U = np.minimum(X1, X2)
V = np.maximum(X1, X2)
print(np.mean(U), np.mean(V), np.cov(U, V)[0, 1])
```

On s'attend à obtenir respectivement des valeurs proches de 1,33, 2,67 et 0,44 (loi faible des grands nombres).