

# Statistiques et Analyse de Données

Paul Lehaut

September 28, 2025

# Contents

<b>1</b>	<b>Rappels</b>	<b>3</b>
1.1	Espérance et (Co)Variance . . . . .	3
1.2	Indépendance . . . . .	3
1.3	Variables Aléatoires Discrètes . . . . .	4
1.4	Variables Aléatoires à Densité . . . . .	4
1.5	Somme de VAs indépendantes . . . . .	4
1.6	Fonction de Répartition . . . . .	4
1.7	Fonction Caractéristique . . . . .	5
1.8	Vecteurs Gaussiens . . . . .	5
1.9	Théorèmes de Convergence . . . . .	5

# 1 Rappels

## 1.1 Espérance et (Co)Variance

On considère dans cette section des  $\text{VAR}$ .

Définition:

L'espérance de  $X \sim \mathbb{P}$  est l'intégral de Lebesgue:

$$\mathbb{E}(X) := \int_{\omega \in \Omega} X(\omega) d\mathbb{P}(\omega) = \int_{x \in \mathbb{R}} x d\mathbb{P}(x)$$

qui est bien définie si  $X$  est intégrable, c'est-à-dire si  $\mathbb{E}(|X|) < +\infty$ .

Si par ailleurs  $\mathbb{E}(X^2) < +\infty$ , alors la variance de  $X$  est bien définie par:

$$\mathbb{V}(X) := \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$

L'inégalité de Jensen donne alors, pour  $X$  une  $\text{VAR}^d$  intégrable, soit  $f$  une fonction à valeurs réelles convexe définie sur  $\mathbb{R}^d$ , alors  $\mathbb{E}(f(X))$  est bien définie et:

$$f(\mathbb{E}(X)) \leq \mathbb{E}(f(X)).$$

Définition:

La covariance entre  $X$  et  $Y$ , telles que  $\mathbb{E}(X) < +\infty$  et  $\mathbb{E}(Y) < +\infty$ , est définie par:

$$\text{Cov}(X, Y) := \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

on a alors l'identité remarquable suivante:

$$\mathbb{V}(X + Y) = \mathbb{V}(X) + 2\text{Cov}(X, Y) + \mathbb{V}(Y).$$

Lorsque  $X$  est une  $\text{VAR}^d$  de carré intégrable, on peut définir sa matrice de covariance qui est symétrique positive.

Définition: Le coefficient de corrélation de  $X$  et  $Y$  est définie par:

$$\rho_{X,Y} := \begin{cases} \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}} \in [-1, 1] & \text{si } \text{Var}(X) \text{Var}(Y) > 0, \\ 0 & \text{sinon.} \end{cases}$$

## 1.2 Indépendance

Définition: Une famille finie de variables aléatoires  $(X_1, \dots, X_n)$  définies sur  $(\Omega, \mathcal{F})$  est indépendante si pour toute famille de sous-ensembles mesurables  $(C_1, \dots, C_n)$  on a:

$$\mathbb{P}(X_1 \in C_1, \dots, X_n \in C_n) = \mathbb{P}(X_1 \in C_1) \dots \mathbb{P}(X_n \in C_n)$$

ou, de façon équivalente, si:

$$\mathbb{P} = \mathbb{P}_{X_1} \otimes \dots \otimes \mathbb{P}_{X_n}.$$

Si  $X_1, \dots, X_n$  sont indépendantes alors, pour toutes fonctions mesurables  $f_1, \dots, f_n$  telles que  $f_1(X_1), \dots, f_n(X_n)$  soient intégrables, on a:

$$\mathbb{E}(f_1(X_1), \dots, f_n(X_n)) = \mathbb{E}(f_1(X_1)) \dots \mathbb{E}(f_n(X_n)).$$

### 1.3 Variables Aléatoires Discrètes

Si  $E$  est discret, toutes mesures de probabilités est caractérisée par la famille de nombres  $(p(x), x \in E)$ . Les variables aléatoires principales sont définies dans le polycopié.

### 1.4 Variables Aléatoires à Densité

Les intégrales sont ici à considérer au sens de Lebesgue.

Définition:

Une variable aléatoire  $X$  est dite à densité  $p$  si:

$$\forall C \in \mathcal{B}(\mathbb{R}^d), \mathbb{P}(X \in C) = \int_{x \in C} p(x) dx$$

une fonction mesurable et positive  $p$  est une densité de probabilité si et seulement si:

$$\int_{x \in \mathbb{R}^d} p(x) dx = 1.$$

Des variables aléatoires  $X_1, \dots, X_n$  sont indépendantes si le vecteur aléatoire  $(X_1, \dots, X_n)$  a pour densité  $p = p_{X_1} \otimes \dots \otimes p_{X_n}$ .

Théorème: Formule de transfert

Si  $X$  a une densité  $p$ , alors pour toutes fonctions mesurables  $f$  telle que  $\mathbb{E}(|f(X)|) < +\infty$ , alors:

$$\mathbb{E}(f(X)) = \int_{x \in \mathbb{R}^d} f(x) p(x) dx.$$

Les exemples principaux de variables aléatoires à densité se trouvent dans le polycopié.

### 1.5 Somme de VAs indépendantes

Soient  $X$  et  $Y$  deux VAs indépendantes, on note  $Z := X + Y$ .

Proposition:

La densité de  $Z$  est:

$$r(z) = \int_{x \in \mathbb{R}^d} p_X(x) p_Y(z - x) dx$$

on l'appelle la convolution de  $p_X$  et  $p_Y$ .

### 1.6 Fonction de Répartition

On se place dans le cas où  $E = \mathbb{R}$ .

Définition:

La fonction de répartition d'un variable aléatoire  $X$  est définie par:

$$\forall x \in \mathbb{R}, F(x) = \mathbb{P}(X \leq x)$$

alors, pour tout réel  $r \in (0, 1)$ , un quantile d'ordre  $r$  pour  $X$  est un nombre  $q_r$  tel que:

$$\mathbb{P}(X \leq q_r) = F(q_r) = r.$$

En général un quantile n'existe pas toujours ou n'est pas unique, néanmoins, lorsque  $X$  possède une densité qui est positive alors le quantile existe et est unique.

## 1.7 Fonction Caractéristique

Soient  $X$  et  $Y$  des vecteurs aléatoires.

Définition: Fonction Caractéristique

La fonction caractéristique de  $X$  est la fonction  $\Psi_X : \mathbb{R}^d \rightarrow \mathbb{C}$  définie par:

$$\Psi_X(u) := \mathbb{E}(e^{i\langle u, X \rangle}) = \mathbb{E}(\cos(\langle u, X \rangle)) + i\mathbb{E}(\sin(\langle u, X \rangle)).$$

Proposition:

Si, pour tout  $u \in \mathbb{R}^d$ ,  $\Phi_X(u) = \Phi_Y(u)$ , alors  $X$  et  $Y$  sont de même loi.

Proposition:

Si  $X \sim \mathcal{N}(\mu, \sigma^2)$  et  $Y \sim \mathcal{N}(\nu, \tau^2)$  sont indépendantes, alors:

$$X + Y \sim \mathcal{N}(\mu + \nu, \sigma^2 + \tau^2).$$

## 1.8 Vecteurs Gaussiens

Soit  $X$  un vecteur aléatoire.

Définition: Vecteur Gaussien

$X$  est un vecteur Gaussien si, pour tout vecteur  $u$ , la variable aléatoire  $\langle u, X \rangle$  est gaussienne.

De ce qui précède, on déduit que, en notant  $\mathbb{E}(X) = m$  et  $Cov(X) = K$ , alors:

$$\langle u, X \rangle \sim \mathcal{N}(\langle u, m \rangle, \langle u, Ku \rangle) \quad \text{et} \quad \Phi_X(u) = \exp(i\langle u, m \rangle - \frac{1}{2}\langle u, Ku \rangle).$$

Par ailleurs, si  $K$  est inversible, alors  $X$  a la densité:

$$\frac{1}{\sqrt{(2\pi)^d \det K}} \exp(-\frac{1}{2}\langle x - m, K^{-1}(x - m) \rangle)$$

sinon  $X$  n'a pas de densité.

## 1.9 Théorèmes de Convergence

Soient  $(X_n)$  et  $X$  des vecteurs aléatoires.

Définition:

$(X_n)$  converge vers  $X$  presque sûrement si:  $\mathbb{P}(\lim X_n = X) = 1$ .

$(X_n)$  converge vers  $X$  en probabilité si, pour tout  $\epsilon > 0$ ,  $\lim \mathbb{P}(\|X_n - X\| \geq \epsilon) = 0$ .

$(X_n)$  converge vers  $X$  en distribution si, pour toute fonction continue et bornée  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ ,  $\mathbb{E}(f(X_n))$  converge vers  $\mathbb{E}(f(X))$ .

Théorème: Convergence Dominée

Supposons que  $(X_n)$  converge vers  $X$  presque sûrement et qu'il existe  $Y$  positive et intégrable telle que:  $\|X_n\| \leq Y$  presque sûrement, alors  $\mathbb{E}(X_n)$  converge vers  $\mathbb{E}(X)$ .

Pour  $f$  une fonction continue, si  $(X_n)$  converge presque sûrement (respectivement en probabilité ou en distribution) vers  $X$  alors  $f(X_n)$  converge presque sûrement vers  $f(X)$  (respectivement en probabilité ou en distribution).

Proposition:

$(X_n)$  converge en distribution vers  $X$  si et seulement si  $\Psi_{X_n}(u)$  converge vers  $\Psi_X(u)$  pour tout  $u$ .

La convergence presque sûre implique la convergence en probabilité qui implique elle-même la convergence en distribution.

Si on se place dans le cadre réel, alors on a les équivalences suivantes:

-  $(X_n)$  converge vers  $X$  en distribution

-  $\mathbb{P}(X_n \leq x) \longrightarrow \mathbb{P}(X \leq x)$  pour tout  $x$  tel que  $\mathbb{P}(X = x) = 0$

-  $\mathbb{P}(X_n < x) \longrightarrow \mathbb{P}(X < x)$  pour tout  $x$  tel que  $\mathbb{P}(X = x) = 0$

On dit que la suite  $(X_n)$  est indépendante et identiquement distribuée si ses variables sont indépendantes et de même loi. On note alors:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

la moyenne empirique de  $X_1, \dots, X_n$ .

Théorème: La Loi Forte des Grands Nombres

Soit  $(X_n)$  une suite de  $\text{VAR}^d$  iid telle que  $\mathbb{E}(\|X_1\|) < +\infty$ , alors:

$$\lim \bar{X}_n = \mathbb{E}(X_1) \text{ presque sûrement.}$$

Théorème: Théorème Central Limite Multivarié

Soit  $(X_n)$  une suite de  $\text{VAR}^d$  iid telle que  $\mathbb{E}(\|X_1\|^2) < +\infty$ , alors:

$$\lim \sqrt{n}(\bar{X}_n - \mathbb{E}(X_1)) = \mathcal{N}(0, \text{Cov}(X_1)) \text{ en distribution.}$$