

# Statistiques et Analyse de Données

Paul Lehaut

October 4, 2025

# Contents

<b>1 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 Charactéristique . . . . .	5
1.8 Vecteurs Gaussiens . . . . .	5
1.9 Théorèmes de Convergence . . . . .	6
<b>2 Estimateur dans un Modèle Paramétrique</b>	<b>7</b>
2.1 Estimateurs . . . . .	7
2.2 Biais et Risque Quadratique . . . . .	7
2.3 Modèle Paramétrique et Estimation du Moment . . . . .	8
2.4 Convergence Normale . . . . .	9

# 1 Rappels

## 1.1 Espérance et (Co)Variance

On considère dans cette section des 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 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 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.

Proposition:

Si  $X$  est une VAR à densité  $p$ , alors sa fonction de répartition est continue et dérivable presque partout, sa dérivée est presque partout égale à sa fonction de densité.

## 1.7 Fonction Charactéristique

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

Définition: Fonction Charactéristique

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:

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

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

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

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 converge en probabilité qui implique elle même la converge en distribution.

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

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

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

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

On dit que la suite  $(X_n)$  est indépendante et indentiquement 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:

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

## 2 Estimateur dans un Modèle Paramétrique

On considère un échantillon  $X_1, \dots, X_n$  de VA iid dans un espace mesurable  $(E, \mathcal{E})$  de loi  $\mathbb{P}$  inconnue qu'on cherche à éclaircir. On cherche par exemple à estimer  $\mathbb{E}_{\mathbb{P}}(X_1)$ ,  $V_{\mathbb{P}}(X_1)$  l'histogramme de  $\mathbb{P}$  ou encore d'autre quantité d'intérêt (QI).

On se restreint à des lois d'une certaine forme caractérisées par un ou certains paramètres. Formellement, on considère une famille de lois:  $\{\mathbb{P}_{\theta}; \theta \in \Theta\}$ .

### 2.1 Estimateurs

On cherche donc à estimer une QI à l'aide de notre échantillon de VA, pour ce faire on va chercher à approcher QI par une fonction de l'échantillon appelée statistique.

Définition:

Une statistique  $T_n$  est une VA de la forme  $T_n = t_n(X_1, \dots, X_n)$  avec  $t_n$  déterministe et qui ne dépend pas de  $\mathbb{P}$ .

On appelle un estimateur une statistique qui vise à approcher une certaine QI.

Définition:

Un estimateur  $Z_n$  d'une QI est dit constant si  $Z_n$  converge en probabilité vers QI, il est fortement constant s'il converge vers QI presque sûrement.

Par exemple, dans le cas où  $E = \mathbb{R}$ , alors la moyenne empirique  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$  est un estimateur fortement constant de  $\mathbb{E}(X_1)$ .

### 2.2 Biais et Risque Quadratique

Le biais et la MSE permettent de quantifier la distance d'un estimateur à sa QI dans un régime non asymptotique (ie n est fini).

Définition:

Soit un estimateur intégrable  $Z_n$ , le biais de  $Z_n$  est:

$$b(Z_n) = \mathbb{E}(Z_n) - QI, \text{ il s'agit de la distance moyenne de l'estimateur à la QI}$$

si ce biais est nul, on dit que l'estimateur est non biaisé.

Il est intéressant de remarquer que la variance empirique est biaisée ce qui motive la définition suivante:

Définition:

Si  $E = \mathbb{R}$ , l'estimateur non-biaisé de la variance est:

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_i)^2.$$

Une mesure plus précise de la distance de l'estimateur à la QI peut être donnée par la MSE.

Définition:

Soit  $Z_n$  un estimateur de carré intégrable, la MSE de  $Z_n$  est définie par:

$$MSE(Z_n) = E(\|Z_n - QI\|^2).$$

Proposition:

On peut a l'égalité suivante:

$$MSE(Z_n) = ||b(Z_n)||^2 + V(Z_n).$$

En général on ne peut pas minimiser à la fois la variance et le biais. En data science il peut être intéressant d'introduire un biais pour réduire la variance du modèle et donc le risque d'overfitting.

### 2.3 Modèle Paramétrique et Estimation du Moment

On s'intéresse ici à l'estimation de la distribution complète  $\mathbb{P}$  de  $X_1$ . Il y a deux approches principales: la méthode non paramétrique (par histogramme) et la méthode paramétrique qui repose sur la supposition que  $\mathbb{P}$  a une certaine forme (comme exponentielle ou gaussienne).

Définition:

Un modèle paramétrique sur  $E$  est un ensemble de mesures de probabilités:

$$\mathcal{P} = \{\mathbb{P}_\theta; \theta \in \Theta\}$$

sur l'espace  $E$ , indexé par un ensemble de paramètres  $\Theta \subset \mathbb{R}^k$ .

Il est important de constater que si, pour deux valeurs distinctes  $\theta$  et  $\theta'$ , on a  $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$  alors il n'est pas possible de distinguer  $\theta$  de  $\theta'$  par simple observation de  $X_1, \dots, X_n$ . On travaillera donc toujours en supposant que la fonction  $\theta \mapsto \mathbb{P}_\theta$  est injective, on dira alors que  $\mathcal{P}$  est identifiable.

On fixe désormais un modèle paramétrique  $\mathcal{P}$ . Pour tout  $\theta$  il est pratique de dénoté par  $\mathbb{P}_\theta$  la mesure de probabilité pour laquelle, pour tout  $n \geq 1$ , les VAs  $X_1, \dots, X_n$  sont iid selon  $\mathbb{P}_\theta$ . On définit de façon similaire  $\mathbb{E}_\theta$ ,  $V_\theta$  etc.

Proposition:

Soit  $X$  une VA $\mathbb{R}$  non déterministe intégrable et qui prend ses valeurs dans un intervalle  $I$ . Soit  $\Phi : I \rightarrow \mathbb{R}$  strictement convexe, alors:  $\Phi(\mathbb{E}(X)) < \mathbb{E}(\Phi(X))$ .

La méthode des moments est une procédure naturelle pour construire des estimateurs, on détaille ici l'exemple pour le modèle exponentielle:  $\{\epsilon(\lambda); \lambda > 0\}$  dans lequel on cherche un estimateur de  $\lambda$ :

- La loi forte des grands nombres donne, pour tout  $\lambda > 0$ :  $\bar{X}_n \rightarrow \mathbb{E}_\lambda(X_1) = \frac{1}{\lambda}$  presque sûrement, donc  $\frac{1}{\bar{X}_n} \rightarrow \lambda$  presque sûrement.  
La continuité de la fonction  $x \mapsto \frac{1}{x}$  sur  $\mathbb{R}_+^*$  assure donc que  $\bar{\lambda}_n = \frac{1}{\bar{X}_n}$  est un estimateur fortement consistant de  $\lambda$ .

La généralisation abstraite de cette méthode s'énonce de la façon suivante:

- Pour l'estimation de  $g(\theta) \in \mathbb{R}^d$ , cela consiste à trouver  $\phi$  et  $m$  des fonctions telles que:

$$\forall \theta \in \Theta, \mathbb{E}_\theta(\phi(X_1)) = m(g(\theta)).$$

Pour le modèle exponentielle on a pris:  $\phi(x) = x$ ,  $g(\lambda) = \lambda$  et  $m(\lambda) = \frac{1}{\lambda}$ .

Alors, la loi forte des grands nombres nous permet donc d'approximer  $m(g(\theta))$  par:  $\frac{1}{n} \sum_{i=1}^n \phi(X_i)$  de sorte que, si  $m$  possède une fonction réciproque continue  $m^{-1}$ , alors:

$$Z_n = m^{-1}\left(\frac{1}{n} \sum_{i=1}^n \phi(X_i)\right)$$

est un estimateur fortement consistant de  $g(\theta)$ .

## 2.4 Convergence Normale

La construction d'un estimateur par la méthode des moments dépend du choix arbitraire de la fonction  $\phi$ , donc différents choix de  $\phi$  peuvent donner différents estimateurs qui sont tous, par construction, fortement consistant. Pour déterminer l'estimateur le plus intéressant en pratique, on peut s'intéresser à celui qui converge 'le plus vite' vers la QI. Cette vitesse de convergence peut être mesurée à l'aide de la notion de variance asymptotique.

### Définition:

Un estimateur consistant  $Z_n$  de  $g(\theta)$  est asymptotiquement normal si, pour tout  $\theta$ , il existe une matrice symétrique positive  $K(\theta) \in \mathcal{M}_d(\mathbb{R})$  telle que:  $\sqrt{n}(Z_n - g(\theta))$  converge en distribution vers  $\mathcal{N}_d(0, K(\theta))$ .

La fonction  $\theta \mapsto K(\theta)$  est appelé la covariance asymptotique de  $Z_n$ .

### Théorème: Méthode delta

Soit  $(\zeta_n)$  une suite de VA à valeurs dans  $\mathbb{R}^k$  et  $a \in \mathbb{R}^k$  telles que  $\zeta_n \rightarrow a$  en probabilité et  $\sqrt{n}(\zeta_n - a)$  converge en distribution vers un vecteur aléatoire  $Y \in \mathbb{R}^k$ . Soit  $\mathcal{U}$  un ouvert de  $\mathbb{R}^k$  qui contient  $a$ , soit  $\Phi : \mathcal{U} \rightarrow \mathbb{R}^d$  de classe  $\mathcal{C}^1$ , alors:

$$\lim_{n \rightarrow +\infty} \sqrt{n}(\Phi(\zeta_n) - \Phi(a)) = \nabla \Phi(a)Y$$

en distribution.

### Théorème: Slutsky

Soit  $((X_n, Y_n))$  une suite de couples de VAs telle que  $(X_n)$  converge en probabilité vers une variable déterministe  $a$  et que  $(Y_n)$  converge en distribution vers une variable aléatoire  $Y$ . Alors  $((X_n, Y_n))$  converge en distribution vers  $(a, Y)$ , et, par conséquent, pour toute fonction continue  $\Psi$ ,  $(\Psi(X_n, Y_n))$  converge en distribution vers  $\Psi(a, Y)$ .