

# Probabilités et Statistiques pour les Sciences des Données

Thomas Letendre

version du 27 avril 2026

# Table des matières

<b>1</b>	<b>Vocabulaire des probabilités</b>	<b>2</b>
1.1	Espace de probabilité . . . . .	2
1.1.1	Tribu . . . . .	2
1.1.2	Mesure de probabilité . . . . .	3
1.2	Variables aléatoires . . . . .	5
1.2.1	Cadre général . . . . .	5
1.2.2	Variables à valeurs dans $\mathbb{R}^n$ . . . . .	5
1.2.3	Fonction de répartition . . . . .	7
1.3	Espérance et moments d'une variable aléatoire réelle . . . . .	7
1.3.1	Espérance . . . . .	7
1.3.2	Variance et moments . . . . .	8
1.4	Indépendance . . . . .	10
1.4.1	Lois marginales et mesure produit . . . . .	10
1.4.2	Évènements et variables indépendantes . . . . .	12
<b>2</b>	<b>Convergence de suites de variables aléatoires</b>	<b>14</b>
2.1	Convergence en probabilité . . . . .	14
2.2	Convergence presque sûre . . . . .	15
2.3	Convergence en loi . . . . .	17
2.3.1	Définition et caractérisations . . . . .	17
2.3.2	Fonction caractéristique . . . . .	18
2.3.3	Théorème de Lévy et théorème central limite . . . . .	19
<b>3</b>	<b>Statistiques</b>	<b>20</b>
3.1	Le point de vue des statistiques . . . . .	20
3.1.1	Exemple introductif . . . . .	20
3.1.2	Vocabulaire des statistiques . . . . .	21
3.2	Estimateurs ponctuels . . . . .	23
3.2.1	Notion d'estimateur . . . . .	23
3.2.2	Qualité d'un estimateur . . . . .	24
3.2.3	Méthode des moments . . . . .	26
3.2.4	Méthode du maximum de vraisemblance . . . . .	27
3.3	Intervalles de confiance . . . . .	29
3.3.1	Intervalles de confiance non-asymptotiques . . . . .	29
3.3.2	Méthode de la fonction pivot . . . . .	30
3.3.3	Retour sur la convergence en loi . . . . .	32
3.3.4	Intervalles de confiance asymptotiques . . . . .	33
3.4	Régression linéaire . . . . .	35
3.4.1	Droite des moindres carrés . . . . .	35
3.4.2	Le modèle linéaire . . . . .	36

# Chapitre 1

## Vocabulaire des probabilités

### 1.1 Espace de probabilité

#### 1.1.1 Tribu

Soit  $\Omega$  un ensemble, on note  $\mathcal{P}(\Omega)$  l'ensemble de ses parties. Pour tout  $A \subset \Omega$  on note  $A^c = \Omega \setminus A$ .

**Définition 1.1.1** (Tribu). Une famille  $\mathcal{A}$  de parties de  $\Omega$  est appelée une *tribu* sur  $\Omega$  si elle vérifie les propriétés suivantes :

1.  $\Omega \in \mathcal{A}$ ;
2. pour tout  $A \in \mathcal{A}$  on a  $A^c \in \mathcal{A}$ ;
3. pour tout suite  $(A_k)_{k \in \mathbb{N}}$  d'éléments de  $\mathcal{A}$  on a  $\bigcup_{k \in \mathbb{N}} A_k \in \mathcal{A}$ .

On dit alors que  $(\Omega, \mathcal{A})$  est un *espace mesurable*.

*Remarque 1.1.2.* Si  $\mathcal{A}$  est une tribu sur  $\Omega$  alors  $\emptyset \in \mathcal{A}$  et  $\mathcal{A}$  est stable par union finie, par intersection finie ou dénombrable, et par différence (si  $A$  et  $B \in \mathcal{A}$  alors  $B \setminus A \in \mathcal{A}$ ).

*Exemple 1.1.3.* •  $\mathcal{P}(\Omega)$  est une tribu sur  $\Omega$ .

- Si  $A \subset \Omega$ , alors  $\{\emptyset, A, A^c, \Omega\}$  est une tribu sur  $\Omega$ .

En général, il est difficile d'expliciter les éléments d'une tribu. On la décrit plutôt par une famille génératrice.

**Lemme 1.1.4.** Soient  $(\mathcal{A}_i)_{i \in I}$  une famille de tribus sur  $\Omega$ , alors  $\mathcal{A} = \bigcap_{i \in I} \mathcal{A}_i$  est une tribu sur  $\Omega$ .

*Démonstration.* Exercice, ou voir cours du premier semestre. □

**Définition 1.1.5** (Tribu engendrée). Soit  $\mathcal{C}$  une famille de parties de  $\Omega$ , l'intersection de toutes les tribus contenant  $\mathcal{C}$  est la plus petite tribu sur  $\Omega$  contenant  $\mathcal{C}$ . Elle est appelée tribu *engendrée* par  $\mathcal{C}$  et on la note  $\sigma(\mathcal{C})$ .

**Définition 1.1.6** (Boréliens). On note  $\mathcal{B}(\mathbb{R}^n)$  la tribu sur  $\mathbb{R}^n$  engendrée par la famille des ouverts. Les éléments de  $\mathcal{B}(\mathbb{R}^n)$  sont appelés *ensembles boréliens*.

Dans la suite, sauf mention explicite du contraire, on munit  $\mathbb{R}^n$  de la tribu  $\mathcal{B}(\mathbb{R}^n)$ . Il existe des parties non boréliennes de  $\mathbb{R}^n$ , mais elles sont difficiles à construire. En pratique, toutes les sous-ensembles de  $\mathbb{R}^n$  que l'on rencontrera seront boréliens, et on ne se posera pas la question.

**Lemme 1.1.7.** La tribu  $\mathcal{B}(\mathbb{R}^n)$  est engendrée par la famille  $\mathcal{F}$  des pavés fermés, i.e. de la forme  $\prod_{i=1}^n [a_i; b_i]$  avec  $a_i \leq b_i$  pour tout  $i \in \llbracket 1, n \rrbracket$ .

*Démonstration.* La tribu  $\mathcal{B}(\mathbb{R}^n)$  contient les ouverts donc les fermés. En particulier  $\mathcal{F} \subset \mathcal{B}(\mathbb{R}^n)$ , et donc  $\sigma(\mathcal{F}) \subset \mathcal{B}(\mathbb{R}^n)$ . Montrons que  $\sigma(\mathcal{F})$  contient les ouverts de  $\mathbb{R}^n$ . On aura alors  $\mathcal{B}(\mathbb{R}^n) \subset \sigma(\mathcal{F})$ , et donc  $\mathcal{B}(\mathbb{R}^n) = \sigma(\mathcal{F})$ .

Soit  $O$  un ouvert de  $\mathbb{R}^n$ , on note  $\mathcal{F}_O$  la famille des pavés inclus dans  $O$  de la forme  $\prod_{i=1}^n [a_i; b_i]$  avec  $a_1, b_1, \dots, a_n, b_n \in \mathbb{Q}$ . Pour tout  $x \in O$ , il existe  $P \in \mathcal{F}_O$  tel que  $x \in P$ , donc  $O = \bigcup_{P \in \mathcal{F}_O} P$ . On a  $\mathcal{F}_O \subset \mathcal{F}$  et  $\mathcal{F}_O$  est dénombrable car  $\mathbb{Q}^{2n}$  l'est. Donc  $O$  est réunion dénombrable d'éléments de  $\mathcal{F}$ , et donc  $O \in \sigma(\mathcal{F})$ . Donc  $\sigma(\mathcal{F})$  contient les ouverts.  $\square$

**Lemme 1.1.8.** *La tribu  $\mathcal{B}(\mathbb{R}^n)$  est engendrée par la famille  $\mathcal{C}$  des pavés de la forme  $\prod_{i=1}^n ]-\infty; x_i]$  avec  $(x_1, \dots, x_n) \in \mathbb{R}^n$ .*

*Démonstration.* Pour tout  $a_1 \leq b_1, \dots, a_n \leq b_n$ , on a

$$\prod_{i=1}^n ]a_i; b_i] = \left( \prod_{i=1}^n ]-\infty; b_i] \right) \setminus \bigcup_{j=1}^n \left( \prod_{i=1}^{j-1} ]-\infty; b_i] \times ]-\infty; a_j] \times \prod_{i=j+1}^n ]-\infty; b_i] \right) \in \sigma(\mathcal{C}),$$

puis  $\prod_{i=1}^n ]a_i; b_i] = \bigcap_{k \in \mathbb{N}} \prod_{i=1}^n ]a_i - \frac{1}{k}; b_i] \in \sigma(\mathcal{C})$ . Donc  $\mathcal{F} \subset \sigma(\mathcal{C})$ , et  $\mathcal{B}(\mathbb{R}^n) = \sigma(\mathcal{F}) \subset \sigma(\mathcal{C})$ .

Inversement, pour tout  $(x_1, \dots, x_n) \in \mathbb{R}^n$ , on a  $\prod_{i=1}^n ]-\infty; x_i] = \bigcup_{k \in \mathbb{N}} \prod_{i=1}^n ]-k; x_i] \in \sigma(\mathcal{F})$ . Donc  $\mathcal{C} \subset \sigma(\mathcal{F}) = \mathcal{B}(\mathbb{R}^n)$  et  $\sigma(\mathcal{C}) \subset \mathcal{B}(\mathbb{R}^n)$ . Finalement,  $\mathcal{B}(\mathbb{R}^n) = \sigma(\mathcal{C})$ .  $\square$

Le même genre d'argument montre que  $\mathcal{B}(\mathbb{R}^n)$  est engendré par les fermés de  $\mathbb{R}^n$ , par les pavés ouverts, par les ensemble de la forme  $\mathbb{R}^{j-1} \times [a_j; b_j] \times \mathbb{R}^{n-j}, \dots$

## 1.1.2 Mesure de probabilité

**Définition 1.1.9** (Mesure de probabilité). Soit  $(\Omega, \mathcal{A})$  un espace mesurable, une *mesure de probabilité* sur  $(\Omega, \mathcal{A})$  est une fonction  $\mathbb{P} : \mathcal{A} \rightarrow [0; 1]$  telle que :

1.  $\mathbb{P}(\Omega) = 1$ ;

2. pour tout suite  $(A_k)_{k \in \mathbb{N}}$  d'éléments de  $\mathcal{A}$  deux à deux disjoints on a  $\mathbb{P}(\bigsqcup_{k \in \mathbb{N}} A_k) = \sum_{k \in \mathbb{N}} \mathbb{P}(A_k)$ .

On dit alors que  $(\Omega, \mathcal{A}, \mathbb{P})$  est un *espace de probabilités*. On appelle  $\Omega$  l'*univers*. Les éléments de  $\mathcal{A}$  sont appelés *événements* et, pour tout  $A \in \mathcal{A}$ , on dit que  $\mathbb{P}(A)$  est la *probabilité* de  $A$ .

**Proposition 1.1.10.** *Soit  $(\Omega, \mathcal{A}, \mathbb{P})$  un espace de probabilités, alors on a les propriétés suivantes.*

- $\mathbb{P}(\emptyset) = 0$ .
- Pour tout  $A, B \in \mathcal{A}$  tels que  $A \subset B$ , on a  $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A)$ . En particulier  $\mathbb{P}(A) \leq \mathbb{P}(B)$  et  $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ .
- Pour tout  $A$  et  $B \in \mathcal{A}$  on a  $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ .
- Pour toute suite croissante  $(A_k)_{k \in \mathbb{N}}$  d'éléments de  $\mathcal{A}$  (i.e. telle que  $A_k \subset A_{k+1}$  pour tout  $k \in \mathbb{N}$ ), on a  $\mathbb{P}(\bigcup_{k \in \mathbb{N}} A_k) = \lim_{k \rightarrow +\infty} \mathbb{P}(A_k)$ .
- Pour toute suite décroissante  $(A_k)_{k \in \mathbb{N}}$  d'éléments de  $\mathcal{A}$  (i.e. telle que  $A_{k+1} \subset A_k$  pour tout  $k \in \mathbb{N}$ ), on a  $\mathbb{P}(\bigcap_{k \in \mathbb{N}} A_k) = \lim_{k \rightarrow +\infty} \mathbb{P}(A_k)$ .

*Démonstration.* Exercice, ou voir cours du premier semestre.  $\square$

*Exemple 1.1.11* (Mesure uniforme sur un ensemble fini). Si  $\Omega$  est fini, la mesure uniforme  $\mathbb{P}$  sur  $\Omega$  est la mesure de probabilité sur  $(\Omega, \mathbb{P}(\Omega))$  telle que  $\mathbb{P}(A) = \frac{\text{Card}(A)}{\text{Card}(\Omega)}$  pour tout  $A \subset \Omega$ .

En particulier, si  $\Omega = \llbracket 1, 6 \rrbracket$ , on peut penser à  $\Omega$  comme l'ensemble des résultats possibles lors du lancer d'un dé. Alors  $\mathbb{P}(A)$  est la probabilité que le résultat appartienne à la partie  $A \subset \Omega$ . Par exemple  $\mathbb{P}(\{2; 4; 6\}) = \frac{1}{2}$  est la probabilité d'obtenir un résultat pair.

*Exemple 1.1.12* (Espace de probabilité discret). Soit  $\Omega$  un ensemble *dénombrable*, c'est-à-dire fini ou en bijection avec  $\mathbb{N}$ . Quitte à numéroter ses éléments, on peut supposer que  $\Omega = \llbracket 1, N \rrbracket$  (si  $\text{Card}(\Omega) = N$ ) ou que  $\Omega = \mathbb{N}$  (si  $\Omega$  est infini). Soit  $(p_\omega)_{\omega \in \Omega}$  une suite à valeurs dans  $[0; 1]$  et telle que  $\sum_{\omega \in \Omega} p_\omega = 1$ . On définit une mesure de probabilité sur  $(\Omega, \mathcal{P}(\Omega))$  par :

$$\forall A \subset \Omega, \quad \mathbb{P}(A) = \sum_{\omega \in A} p_\omega = \sum_{\omega \in \Omega} p_\omega \mathbf{1}_A(\omega),$$

où  $\mathbf{1}_A : \Omega \rightarrow \{0; 1\}$  est la fonction indicatrice de  $A$ . Inversement, si  $\mathbb{P}$  est une mesure de probabilité sur  $(\Omega, \mathcal{P}(\Omega))$  elle est de la forme précédente avec  $p_\omega = \mathbb{P}(\{\omega\})$  pour tout  $\omega \in \Omega$ .

Tout espace de probabilité de la forme  $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$  avec  $\Omega$  dénombrable est dit *discret*. La fonction  $p : \omega \mapsto \mathbb{P}(\{\omega\}) = p_\omega$  (ou de façon équivalente la suite  $(p_\omega)_{\omega \in \Omega}$ ) est appelée la *fonction de probabilité* associée à  $\mathbb{P}$ .

En général, construire de bonnes mesures de probabilités est un problème difficile. Par exemple, pour des raisons qui sortent du cadre de ce cours, il est impossible de définir une probabilité  $\mathbb{P}$  sur  $([0; 1], \mathcal{P}([0; 1]))$  telle que  $\mathbb{P}([a; b]) = b - a$  pour tout  $a$  et  $b \in [0; 1]$ . C'est pour contourner ce genre de problème qu'on considère des tribus plus petites que l'ensemble des parties, comme les boréliens.

**Théorème 1.1.13** (Carathéodory). *Soit  $f : \mathbb{R}^n \rightarrow [0, +\infty[$  une fonction intégrable et telle que  $\int_{\mathbb{R}^n} f(x) dx = 1$ , alors il existe une unique mesure de probabilité sur  $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$  telle que, pour tout  $a_1 \leq b_1, \dots, a_n \leq b_n$  on a :*

$$\mathbb{P}\left(\prod_{i=1}^n [a_i; b_i]\right) = \int_{\prod_{i=1}^n [a_i; b_i]} f(x_1, \dots, x_n) dx_1 \dots dx_n.$$

*Démonstration.* Difficile et admis. Voir [3, app. A]. □

**Définition 1.1.14** (Densité). Une mesure  $\mathbb{P}$  du type précédent est appelée *mesure à densité*. On dit que  $f$  est la *densité* de  $\mathbb{P}$ .

*Exemple 1.1.15* (Mesure uniforme sur  $[a; b]$ ). Soient  $a < b$ , alors  $f = \frac{1}{b-a} \mathbf{1}_{[a; b]}$  est intégrable sur  $\mathbb{R}$  et d'intégrale 1. La mesure de probabilité  $\mathbb{P}$  sur  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$  de densité  $f$  est dite *uniforme* sur  $[a; b]$ . Comme  $\mathbb{P}([a; b]) = 1$ , on la considère souvent comme une mesure de probabilité sur  $[a; b]$  muni de sa tribu borélienne  $\mathcal{B}([a; b]) = \{[a; b] \cap B \mid B \in \mathcal{B}(\mathbb{R}^n)\}$ .

Dans le théorème de Carathéodory, on peut déduire l'unicité du résultat suivant.

**Théorème 1.1.16** (Lemme de classe monotone). *Soit  $\Omega$  un ensemble et  $\mathcal{C}$  une famille de parties de  $\Omega$  stable par intersection finie (si  $B$  et  $C \in \mathcal{C}$  alors  $B \cap C \in \mathcal{C}$ ). Soient  $\mathbb{P}_1$  et  $\mathbb{P}_2$  deux mesures de probabilités sur  $(\Omega, \sigma(\mathcal{C}))$ . Si  $\mathbb{P}_1(C) = \mathbb{P}_2(C)$  pour tout  $C \in \mathcal{C}$ , alors  $\mathbb{P}_1 = \mathbb{P}_2$ .*

*En particulier, deux mesures de probabilités sur  $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$  qui coïncident sur la famille  $\mathcal{F}$  des pavés fermés sont égales. De même, si elles coïncident sur les pavés de la forme  $\prod_{i=1}^n ]-\infty, x_i]$ , avec  $(x_1, \dots, x_n) \in \mathbb{R}^n$ .*

*Démonstration.* Moins difficile mais admis quand même. Voir [3, app. A]. □

**Définition 1.1.17** (Évènement presque sûr). Soit  $(\Omega, \mathcal{A}, \mathbb{P})$  un espace de probabilité, un évènement  $A \in \mathcal{A}$  est dit *presque sûr* (p.s.) si  $\mathbb{P}(A) = 1$ .

*Exemple 1.1.18.* On considère la mesure de probabilité uniforme sur  $[0; 1]$ . Alors  $\mathbb{P}([0; 1]) = \int_0^1 dt = 1$ , donc  $[0; 1]$  est un évènement presque sûr. Pour tout  $x \in [0; 1]$ , on a  $\mathbb{P}(\{x\}) = \int_x^x dt = 0$ . Donc  $\mathbb{P}([0; 1] \cap \mathbb{Q}) = \mathbb{P}\left(\bigsqcup_{x \in [0; 1] \cap \mathbb{Q}} \{x\}\right) = \sum_{x \in [0; 1] \cap \mathbb{Q}} \mathbb{P}(\{x\}) = 0$ , et donc  $\mathbb{P}([0; 1] \setminus \mathbb{Q}) = 1$  et  $[0; 1] \setminus \mathbb{Q}$  est aussi un évènement presque sûr.

## 1.2 Variables aléatoires

### 1.2.1 Cadre général

**Définition 1.2.1** (Variable aléatoire). Soient  $(\Omega, \mathcal{A}, \mathbb{P})$  un espace de probabilité et  $(E, \mathcal{B})$  un espace mesurable. Une fonction  $X : \Omega \rightarrow E$  est appelée une *variable aléatoire* (v.a.) si elle est *mesurable*, i.e. si pour tout  $B \in \mathcal{B}$  on a  $X^{-1}(B) \in \mathcal{A}$ . Pour tout  $B \in \mathcal{B}$ , on note alors  $(X \in B)$  l'évènement  $X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\}$ .

*Remarque 1.2.2.* Si  $(\Omega, \mathcal{A}, \mathbb{P})$  est discret (i.e. si  $\mathcal{A} = \mathcal{P}(\Omega)$ ) alors toute fonction de  $\Omega$  vers un espace mesurable est une variable aléatoire.

**Lemme 1.2.3.** Si  $X$  est une telle variable aléatoire, alors on définit une mesure de probabilité  $\mathbb{P}_X$  sur  $(E, \mathcal{B})$  par  $\mathbb{P}_X : B \mapsto \mathbb{P}(X^{-1}(B)) = \mathbb{P}(X \in B)$ .

*Démonstration.* On a  $\mathbb{P}_X(E) = \mathbb{P}(X^{-1}(E)) = \mathbb{P}(\Omega) = 1$ . Par ailleurs, si  $(B_k)_{k \in \mathbb{N}}$  sont des éléments deux à deux disjoints de  $\mathcal{B}$ , alors les  $(X^{-1}(B_k))_{k \in \mathbb{N}}$  sont des éléments deux à deux disjoints de  $\mathcal{A}$  et

$$\mathbb{P}_X\left(\bigsqcup_{k \in \mathbb{N}} B_k\right) = \mathbb{P}\left(X^{-1}\left(\bigsqcup_{k \in \mathbb{N}} B_k\right)\right) = \mathbb{P}\left(\bigsqcup_{k \in \mathbb{N}} X^{-1}(B_k)\right) = \sum_{k \in \mathbb{N}} \mathbb{P}(X^{-1}(B_k)) = \sum_{k \in \mathbb{N}} \mathbb{P}_X(B_k). \quad \square$$

**Définition 1.2.4** (Loi d'une variable aléatoire). La mesure  $\mathbb{P}_X$  est appelée la *loi* ou la *distribution* de la variable aléatoire  $X$ .

*Remarque 1.2.5.* • Toute mesure de probabilité est la loi d'une variable aléatoire. En effet, si  $\mathbb{P}$  est une probabilité sur  $(\Omega, \mathcal{A})$ , alors  $\mathbb{P} = \mathbb{P}_X$  où  $X = \text{Id}_{|\Omega}$ .

- Souvent, on ne s'intéresse qu'à la loi de  $X$ . Dans ce cas, on peut directement travailler avec  $(E, \mathcal{B}, \mathbb{P}_X)$ , sans avoir à expliciter  $(\Omega, \mathcal{A}, \mathbb{P})$ .
- Si, par exemple,  $X$  est une variable de Bernoulli de paramètre  $\frac{1}{2}$ , c'est-à-dire à valeurs dans  $\{0; 1\}$  et  $\mathbb{P}(X = 0) = \frac{1}{2} = \mathbb{P}(X = 1)$ , alors  $X$  et  $1 - X$  sont deux variables aléatoires différentes de même loi. Si on veut les différencier, on ne peut pas juste regarder leurs lois. Il faut les considérer en tant que fonctions, et donc connaître  $(\Omega, \mathcal{A}, \mathbb{P})$ .

**Lemme 1.2.6.** Soit  $X : \Omega \rightarrow E$  où  $(\Omega, \mathcal{A}, \mathbb{P})$  est un espace de probabilité et  $(E, \mathcal{B})$  est un espace mesurable. Soit  $\mathcal{C} \subset \mathcal{B}$  telle que  $\sigma(\mathcal{C}) = \mathcal{B}$ , alors  $X$  est une variable aléatoire si et seulement si, pour tout  $B \in \mathcal{C}$ ,  $X^{-1}(B) \in \mathcal{A}$ .

*Démonstration.* La condition est nécessaire puisque  $\mathcal{C} \subset \mathcal{B}$ . Soit  $\mathcal{B}' = \{B \subset E \mid X^{-1}(B) \in \mathcal{A}\}$ , on a  $\mathcal{C} \subset \mathcal{B}'$ . Montrons que  $\mathcal{B}'$  est une tribu. On aura alors  $\mathcal{B} = \sigma(\mathcal{C}) \subset \mathcal{B}'$ , ce qui prouvera le résultat.

On a  $X^{-1}(E) = \Omega$ , donc  $E \in \mathcal{B}'$ . Si  $B \in \mathcal{B}'$  alors  $X^{-1}(E \setminus B) = \Omega \setminus X^{-1}(B)$ . Comme  $X^{-1}(B)$  est dans  $\mathcal{A}$  son complémentaire aussi, et donc  $B^c \in \mathcal{B}'$ . Enfin, soit  $(B_k)_{k \in \mathbb{N}}$  des éléments de  $\mathcal{B}'$ , on a  $X^{-1}(\bigcup_{k \in \mathbb{N}} B_k) = \bigcup_{k \in \mathbb{N}} X^{-1}(B_k) \in \mathcal{A}$  et donc  $\bigcup_{k \in \mathbb{N}} B_k \in \mathcal{B}'$ .  $\square$

### 1.2.2 Variables à valeurs dans $\mathbb{R}^n$

Dorénavant, on considère des variables aléatoires à valeurs dans  $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ . On parlera de *variable aléatoire réelle* (v.a.r.) si  $n = 1$  et de *vecteur aléatoire* si  $n \geq 2$ . D'après le lemme 1.2.6, pour prouver que  $X$  est une variable aléatoire de  $(\Omega, \mathcal{A}, \mathbb{P})$  dans  $\mathbb{R}^n$ , il suffit de montrer que  $X^{-1}(B) \in \mathcal{A}$  pour tout  $B$  dans l'une des familles suivantes : les ouverts de  $\mathbb{R}^n$  ; les fermés de  $\mathbb{R}^n$  ; les pavés ouverts ; les pavés fermés ; les pavés de la forme  $\prod_{i=1}^n ]-\infty; x_i]$  avec  $(x_1, \dots, x_n) \in \mathbb{R}^n$ .

**Corollaire 1.2.7.** Soient  $X : \Omega \rightarrow \mathbb{R}^n$  une variable aléatoire et  $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}^m$  une fonction continue, alors  $\varphi(X)$  est une variable aléatoire à valeurs dans  $\mathbb{R}^m$ .

*Démonstration.* Soit  $O \subset \mathbb{R}^m$  un ouvert. Par continuité  $\varphi^{-1}(O)$  est ouvert de  $\mathbb{R}^n$  donc borélien. Donc  $(\varphi \circ X)^{-1}(O) = X^{-1}(\varphi^{-1}(O)) \in \mathcal{A}$ . Donc, pour tout  $O$  ouvert de  $\mathbb{R}^m$ , on a  $(\varphi \circ X)^{-1}(O) \in \mathcal{A}$ .  $\square$

**Corollaire 1.2.8.** Soit  $X = (X_1, X_2, \dots, X_n)$  de  $\Omega$  dans  $\mathbb{R}^n$ , alors  $X$  est un vecteur aléatoire si et seulement si les  $(X_i)_{1 \leq i \leq n}$  sont des variables aléatoires réelles.

*Démonstration.* Notons  $\pi_i : (x_1, \dots, x_n) \mapsto x_i$  de  $\mathbb{R}^n$  dans  $\mathbb{R}$ . Si  $X$  est un vecteur aléatoire alors, pour tout  $i \in \llbracket 1, n \rrbracket$ ,  $X_i = \pi_i \circ X$  est une v.a.r. par continuité de  $\pi_i$ . Inversement, si les  $(X_i)_{1 \leq i \leq n}$  sont des v.a.r., pour tout  $a_1 \leq b_1, \dots, a_n \leq b_n$  on a :

$$X^{-1}\left(\prod_{i=1}^n [a_i; b_i]\right) = \{\omega \in \Omega \mid \forall i \in \llbracket 1, n \rrbracket, X_i(\omega) \in [a_i; b_i]\} = \bigcap_{i=1}^n X_i^{-1}([a_i; b_i]) \in \mathcal{A}.$$

Donc  $X$  est un vecteur aléatoire d'après les lemmes 1.2.6 et 1.1.7.  $\square$

**Définition 1.2.9** (Variable discrète). On dit qu'une variable aléatoire  $X : \Omega \rightarrow \mathbb{R}^n$  est *discrète* si  $X(\Omega)$  est un ensemble dénombrable.

**Lemme 1.2.10.** Soient  $D \subset \mathbb{R}^n$  un ensemble dénombrable et  $X : \Omega \rightarrow D$ . La fonction  $X$  est une variable aléatoire discrète sur  $(\Omega, \mathcal{A}, \mathbb{P})$  si et seulement si, pour tout  $x \in D$ , on a  $X^{-1}(\{x\}) \in \mathcal{A}$ .

*Démonstration.* Si  $X$  est une variable aléatoire alors elle est discrète. De plus, pour tout  $x \in D$  on a  $\{x\} \in \mathcal{B}(\mathbb{R}^n)$  et donc  $X^{-1}(\{x\}) \in \mathcal{A}$ .

Inversement, supposons cette condition réalisée. Soit  $B \in \mathcal{B}(\mathbb{R}^n)$ , alors  $B \cap D$  est dénombrable. Donc  $X^{-1}(B) = X^{-1}(B \cap D) = \bigsqcup_{x \in B \cap D} X^{-1}(\{x\}) \in \mathcal{A}$ . C'est vrai pour tout  $B \in \mathcal{B}(\mathbb{R}^n)$ , donc  $X$  est bien une variable aléatoire.  $\square$

**Corollaire 1.2.11.** Soient  $X$  une v.a. discrète à valeurs dans  $D \subset \mathbb{R}^n$  et  $\varphi : D \rightarrow \mathbb{R}^m$ , alors  $\varphi(X)$  est une v.a. discrète.

*Démonstration.* On a  $(\varphi \circ X)(\Omega) \subset \varphi(D)$  donc c'est un ensemble dénombrable. Soit  $y \in \varphi(D)$  alors  $\varphi^{-1}(\{y\}) \subset D$  est dénombrable et

$$(\varphi \circ X)^{-1}(\{y\}) = X^{-1}(\varphi^{-1}(\{y\})) = X^{-1}\left(\bigsqcup_{x \in \varphi^{-1}(\{y\})} \{x\}\right) = \bigsqcup_{x \in \varphi^{-1}(\{y\})} X^{-1}(\{x\}) \in \mathcal{A}.$$

Donc  $(\varphi \circ X)^{-1}(\{y\}) \in \mathcal{A}$  pour tout  $y \in \varphi(D)$  et  $\varphi(X)$  est une v.a. discrète par le lemme 1.2.10.  $\square$

**Lemme 1.2.12.** Soient  $X : \Omega \rightarrow \mathbb{R}^n$  une variable aléatoire discrète et  $D = X(\Omega) \subset \mathbb{R}^n$ . La loi  $\mathbb{P}_X$  de  $X$  est totalement déterminée par sa fonction de probabilité, i.e. par  $(\mathbb{P}(X = x))_{x \in D}$ .

*Démonstration.* Pour tout  $B \in \mathcal{B}(\mathbb{R}^n)$ , comme  $B \cap D \subset D$  est dénombrable, on a

$$\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}(X^{-1}(B \cap D)) = \mathbb{P}\left(\bigsqcup_{x \in B \cap D} X^{-1}(\{x\})\right) = \sum_{x \in B \cap D} \mathbb{P}(X = x). \quad \square$$

**Définition 1.2.13** (Variable à densité). On dit qu'une variable aléatoire  $X : \Omega \rightarrow \mathbb{R}^n$  est à *densité* si sa loi  $\mathbb{P}_X$  est une mesure à densité, i.e. s'il existe  $f : \mathbb{R}^n \rightarrow [0, +\infty[$  intégrable d'intégrale 1 telle que, pour tout  $a_1 \leq b_1, \dots, a_n \leq b_n$ , on a  $\mathbb{P}(X \in \prod_{i=1}^n [a_i; b_i]) = \int_{\prod_{i=1}^n [a_i; b_i]} f(x_1, \dots, x_n) dx_1 \dots dx_n$ .

**Lemme 1.2.14.** Soient  $X : \Omega \rightarrow \mathbb{R}^n$  une variable aléatoire admettant une densité  $f$ . La loi  $\mathbb{P}_X$  de  $X$  est totalement déterminée par  $f$ .

*Démonstration.* La densité  $f$  détermine  $\mathbb{P}_X$  sur la famille  $\mathcal{F}$  des pavés fermés, qui est stable par intersection et engendre  $\mathcal{B}(\mathbb{R}^n)$ , voir lemme 1.1.7. Donc  $f$  détermine totalement  $\mathbb{P}_X$  d'après le lemme de classe monotone, voir théorème 1.1.16.  $\square$

### 1.2.3 Fonction de répartition

**Définition 1.2.15** (Fonction de répartition). Soit  $X$  une variable aléatoire à valeurs dans  $\mathbb{R}^n$ . Sa *fonction de répartition* est la fonction  $F_X : (x_1, \dots, x_n) \mapsto \mathbb{P}(X \in \prod_{i=1}^n ]-\infty; x_i])$  de  $\mathbb{R}^n$  dans  $[0; 1]$ .

**Lemme 1.2.16.** La fonction  $F_X$  caractérise la loi de  $X$ .

*Démonstration.* Si on connaît  $F_X$  alors on connaît  $\mathbb{P}_X(\prod_{i=1}^n ]-\infty; x_i]) = \mathbb{P}(X \in \prod_{i=1}^n ]-\infty; x_i])$  pour tout  $(x_1, \dots, x_n) \in \mathbb{R}^n$ . Comme la famille  $\mathcal{C}$  des pavés de cette forme est stable par intersection et engendre  $\mathcal{B}(\mathbb{R}^n)$ , voir lemme 1.1.8, on déduit le résultat du lemme de classe monotone 1.1.16.  $\square$

**Proposition 1.2.17.** Soit  $X$  une v.a.r., alors  $F_X : \mathbb{R} \rightarrow [0; 1]$  a les propriétés suivantes.

- $F_X$  est croissante, en particulier elle admet une limite à droite et à gauche en tout point.
- $F_X$  est continue à droite : pour tout  $x_0 \in \mathbb{R}$ ,  $F_X(x_0) \xrightarrow{x \rightarrow x_0^+} F_X(x_0)$ .
- $F_X(x) \xrightarrow{x \rightarrow -\infty} 0$  et  $F_X(x) \xrightarrow{x \rightarrow +\infty} 1$ .
- Pour tout  $a \leq b$  on a  $P_X([a; b]) = F_X(b) - \lim_{x \rightarrow a^-} F_X(x)$ .

*Démonstration.* Exercice en utilisant les propriétés de  $\mathbb{P}_X$ , voir proposition 1.1.10.  $\square$

*Exemple 1.2.18.* Soit  $X$  une v.a.r. discrète, disons à valeur dans  $\mathbb{Z}$ . Pour tout  $n \in \mathbb{Z}$ , la fonction  $F_X$  est constante sur  $[n; n + 1[$ , égale à  $\mathbb{P}_X(]-\infty; n]) = \sum_{k \leq n} \mathbb{P}(X = k)$ . En  $n + 1$ , la fonction présente un saut de hauteur  $\mathbb{P}(X = n + 1)$ .

*Exemple 1.2.19.* Soit  $X$  une v.a.r. admettant une densité  $f$ . Alors  $F_X : x \mapsto \int_{-\infty}^x f(t) dt$  est continue. De plus, si  $f$  est continue en  $x \in \mathbb{R}$ , alors  $F_X$  est dérivable en  $x$  et  $F'_X(x) = f(x)$ .

## 1.3 Espérance et moments d'une variable aléatoire réelle

Dans cette section, on fixe un espace de probabilité abstrait  $(\Omega, \mathcal{A}, \mathbb{P})$  sur lequel seront définies toutes nos variables aléatoires.

### 1.3.1 Espérance

**Définition 1.3.1** (Espérance). Soit  $X$  une v.a. réelle. Si  $X$  est à valeurs dans  $[0; +\infty[$ , on peut définir son *espérance*  $\mathbb{E}[X] \in [0; +\infty]$ . Si  $X$  est à valeurs dans  $\mathbb{R}$  et  $\mathbb{E}[|X|] < +\infty$ , on dit que  $X$  est *intégrable* ou  $L^1$ , et on peut définir son *espérance*  $\mathbb{E}[X] \in \mathbb{R}$ . Dans les deux cas,  $\mathbb{E}[X]$  est défini par :

- $\mathbb{E}[X] = \sum_{x \in D} x \mathbb{P}(X = x)$ , si  $X$  est discrète à valeurs dans  $D$  dénombrable ;
- $\mathbb{E}[X] = \int_{\mathbb{R}} x f(x) dx$ , si  $X$  admet une densité  $f$  ;
- $\mathbb{E}[X] = \int_{\Omega} X(\omega) d\mathbb{P}(\omega)$ , en général.

On dit qu'une v.a.r. intégrable  $X$  est *centrée* si  $\mathbb{E}[X] = 0$ .

Dans le dernier cas, on a besoin de la théorie de la mesure pour donner du sens à  $\mathbb{E}[X]$ , ce qui sort du cadre de ce cours. On se contentera donc de retenir qu'on peut donner du sens à  $\mathbb{E}[X]$ , et que cette quantité s'interprète comme la moyenne des valeurs de  $X$  sur  $\Omega$ , pour la répartition de masse donnée par  $\mathbb{P}$ . Si  $X$  est une v.a.r. positive ou  $L^1$ , on peut prouver que  $\mathbb{E}[X] = \int_{\mathbb{R}} x d\mathbb{P}_X(x)$  (avec de la théorie de la mesure toujours). En particulier,  $\mathbb{E}[X]$  ne dépend que de la loi  $\mathbb{P}_X$  de  $X$ .

*Exemple 1.3.2.* Soit  $A \in \mathcal{A}$ , sa fonction indicatrice  $\mathbf{1}_A : \Omega \rightarrow \{0; 1\}$  définit une v.a.r. discrète et  $\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(\mathbf{1}_A(\omega) = 1) = \mathbb{P}(A)$ .

**Proposition 1.3.3** (Propriétés de l'espérance). *Soient  $X$  et  $Y$  des v.a.r. et  $\lambda \in \mathbb{R}$ .*

- *Si  $X \geq Y \geq 0$  alors  $\mathbb{E}[X] \geq \mathbb{E}[Y] \geq 0$ . De plus  $\mathbb{E}[X] = 0$  si et seulement si  $X = 0$  p.s.*
- *Si  $\lambda \geq 0$  et  $X$  et  $Y$  sont à valeurs positives alors  $\mathbb{E}[X + \lambda Y] = \mathbb{E}[X] + \lambda \mathbb{E}[Y]$ .*
- *Si  $X$  et  $Y$  sont  $L^1$  alors  $X + \lambda Y$  est  $L^1$  et  $\mathbb{E}[X + \lambda Y] = \mathbb{E}[X] + \lambda \mathbb{E}[Y]$ .*

*Démonstration.* Admis. □

**Définition 1.3.4** (Espace  $L^1$ ). On note  $L^1(\Omega, \mathcal{A}, \mathbb{P})$  l'ensemble des v.a.r. intégrables sur  $(\Omega, \mathcal{A}, \mathbb{P})$ . La notation est parfois abrégée en  $L^1(\Omega)$  si  $\mathcal{A}$  et  $\mathbb{P}$  sont clairs d'après le contexte.

*Remarque 1.3.5.* Le dernier point de la proposition 1.3.3 affirme que  $L^1(\Omega)$  est un  $\mathbb{R}$ -espace vectoriel, et que  $\mathbb{E} : L^1(\Omega) \rightarrow \mathbb{R}$  est une forme linéaire sur cet espace.

**Théorème 1.3.6** (De transfert). *Soient  $X$  une v.a. à valeurs dans  $\mathbb{R}^n$  et  $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}$ .*

- *Si  $X$  est discrète, à valeurs dans  $D \subset \mathbb{R}^n$  dénombrable, et si  $\sum_{x \in D} |\varphi(x)| \mathbb{P}(X = x) < +\infty$ , alors  $\varphi(X)$  est une v.a.r. intégrable et  $\mathbb{E}[\varphi(X)] = \sum_{x \in D} \varphi(x) \mathbb{P}(X = x)$ .*
- *Si  $X$  admet la densité  $f$  et si  $|\varphi|f$  est intégrable sur  $\mathbb{R}^n$ , alors  $\varphi(X)$  est une v.a.r. intégrable et  $\mathbb{E}[\varphi(X)] = \int_{\mathbb{R}^n} \varphi(x) f(x) dx$ .*

*Démonstration.* Admis. □

**Proposition 1.3.7** (Markov). *Soit  $X$  une v.a.r.  $L^1$ . Pour tout  $C > 0$ , on a  $\mathbb{P}(X \geq C) \leq \frac{\mathbb{E}[X]}{C}$ .*

*Démonstration.* Notons  $A = \{\omega \in \Omega \mid X(\omega) \geq C\} = X^{-1}([C; +\infty]) \in \mathcal{A}$ . Pour tout  $\omega \in \Omega$ , on a  $C\mathbf{1}_A(\omega) \leq X(\omega)$ . Donc  $X - C\mathbf{1}_A$  est à valeurs positive et  $\mathbb{E}[X - C\mathbf{1}_A] \geq 0$ . Donc

$$\mathbb{E}[X] \geq C\mathbb{E}[\mathbf{1}_A] = C\mathbb{P}(A) = C\mathbb{P}(X \geq C). \quad \square$$

## 1.3.2 Variance et moments

**Définition 1.3.8** (Variable  $L^p$ ). Soient  $X$  une v.a.r. et  $p \in \mathbb{N}^*$ , on dit que  $X$  est  $L^p$  ou que  $X$  admet un moment d'ordre  $p$  si  $\mathbb{E}[|X|^p] < +\infty$ . On dit que  $X$  est  $L^\infty$  ou bornée presque sûrement s'il existe  $M \geq 0$  tel que  $\mathbb{P}(|X| \leq M) = 1$ .

**Définition 1.3.9** (Espace  $L^p$ ). Soit  $p \in \mathbb{N}^* \sqcup \{\infty\}$ , on note  $L^p(\Omega, \mathcal{A}, \mathbb{P})$  (ou parfois  $L^p(\Omega)$ ) l'ensemble des v.a.r.  $L^p$  sur  $(\Omega, \mathcal{A}, \mathbb{P})$ .

**Lemme 1.3.10.** *Pour tout  $p \in \mathbb{N}^* \sqcup \{\infty\}$ , l'ensemble  $L^p(\Omega, \mathcal{A}, \mathbb{P})$  est un  $\mathbb{R}$ -espace vectoriel.*

*Démonstration.* Montrons que  $L^p(\Omega)$  est un sous-ev de l'espace des fonctions de  $\Omega$  dans  $\mathbb{R}$ . On traite d'abord le cas  $p = \infty$ . Soient  $X$  et  $Y \in L^\infty(\Omega)$  et  $\lambda \in \mathbb{R}$ . Il existe  $M \geq 0$  et  $N \geq 0$  tels que

$\mathbb{P}(|X| \leq M) = 1 = \mathbb{P}(|Y| \leq N)$ . Si  $|X| \leq M$  et  $|Y| \leq N$  alors  $|X + \lambda Y| \leq |X| + |\lambda||Y| \leq M + |\lambda|N$ . Par contraposée, si  $|X + \lambda Y| > M + |\lambda|N$  alors  $|X| > M$  ou  $|Y| > N$ . Donc, en notant  $C = M + |\lambda|N$ ,

$$\mathbb{P}(|X + \lambda Y| > C) \leq \mathbb{P}(|X| > M \text{ ou } |Y| > N) \leq \mathbb{P}(|X| > M) + \mathbb{P}(|Y| > N) = 0.$$

Donc  $\mathbb{P}(|X + \lambda Y| \leq C) = 1$  et  $X + \lambda Y \in L^\infty(\Omega)$ .

Soit  $p \in \mathbb{N}^*$ , soient  $X$  et  $Y \in L^p(\Omega)$  et  $\lambda \in \mathbb{R}$ , on a

$$|X + \lambda Y|^p \leq (|X| + |\lambda||Y|)^p = \sum_{k=0}^p \binom{p}{k} |\lambda|^k \underbrace{|Y|^k |X|^{p-k}}_{\leq |X|^p + |Y|^p} \leq (|X|^p + |Y|^p)(1 + |\lambda|)^p.$$

Donc  $\mathbb{E}[|X + \lambda Y|^p] \leq (1 + |\lambda|)^p(\mathbb{E}[|X|^p] + \mathbb{E}[|Y|^p]) < +\infty$  et  $X + \lambda Y \in L^p(\Omega)$ .  $\square$

**Lemme 1.3.11.** Soient  $p$  et  $q \in \mathbb{N}^* \sqcup \{\infty\}$  tels que  $p \geq q$ , alors  $L^p(\Omega) \subset L^q(\Omega)$ .

*Démonstration.* Soit  $X \in L^p(\Omega)$ . Supposons d'abord  $p \in \mathbb{N}^*$ . Pour tout  $\omega \in \Omega$ , si  $|X(\omega)| \leq 1$  on a  $|X(\omega)|^q \leq 1$ ; si  $|X(\omega)| \geq 1$  on a  $|X(\omega)|^q \leq |X(\omega)|^p$ . Dans tous les cas  $|X(\omega)|^q \leq 1 + |X(\omega)|^p$ . Donc  $\mathbb{E}[|X|^q] \leq \mathbb{E}[1 + |X|^p] \leq 1 + \mathbb{E}[|X|^p] < +\infty$  et  $X \in L^q(\Omega)$ .

Si maintenant  $p = \infty$ , il existe  $M \geq 0$  tel que  $A = \{\omega \in \Omega \mid |X(\omega)| \leq M\}$  est presque sûr. On peut supposer  $q < p$ , et alors  $\mathbf{1}_A |X|^q \leq M^q$ . Comme  $A^c$  est de proba nulle, on a  $\mathbf{1}_{A^c} |X|^q$  positive et nulle p.s., donc  $\mathbb{E}[|X|^q] = \mathbb{E}[\mathbf{1}_A |X|^q] + \mathbb{E}[\mathbf{1}_{A^c} |X|^q] = \mathbb{E}[\mathbf{1}_A |X|^q] \leq M^q < +\infty$  et  $X \in L^q(\Omega)$ .  $\square$

*Remarque 1.3.12.* La preuve montre que si  $|X| \leq M$  presque sûrement alors  $\mathbb{E}[|X|] \leq M$ .

**Définition 1.3.13** (Moments). Soient  $p \in \mathbb{N}^*$  et  $X \in L^p(\Omega)$ , on appelle *moment* (resp. *moment centré*) d'ordre  $p$  de  $X$  la quantité  $\mathbb{E}[X^p]$  (resp.  $\mathbb{E}[(X - \mathbb{E}[X])^p]$ ).

**Définition 1.3.14** (Variance). Soit  $X \in L^2(\Omega)$ , sa *variance* est son moment centré d'ordre 2 :

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

L'*écart-type* de  $X$  est  $\sqrt{\text{Var}(X)}$ . On dit que  $X$  est *réduite* si  $\text{Var}(X) = 1$ .

**Proposition 1.3.15** (Propriétés de la variance). Soit  $X \in L^2(\Omega)$ , on a :

- $\text{Var}(X) = 0$  si et seulement si  $X = \mathbb{E}[X]$  p.s., c'est-à-dire  $X$  est constante p.s.
- pour tout  $a$  et  $b \in \mathbb{R}$ ,  $\text{Var}(aX + b) = a^2 \text{Var}(X)$ .

*Démonstration.* Exercice.  $\square$

**Lemme 1.3.16** (Inégalité de concentration). Soient  $p \in \mathbb{N}^*$  et  $X \in L^p(\Omega)$ , pour tout  $\varepsilon > 0$  on a

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \frac{1}{\varepsilon^p} \mathbb{E}[|X - \mathbb{E}[X]|^p].$$

En particulier, si  $X \in L^2(\Omega)$ , pour tout  $\varepsilon > 0$  on a  $\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \frac{\text{Var}(X)}{\varepsilon^2}$ . C'est l'inégalité de Bienaymé–Chebyshev.

*Démonstration.* Soit  $\varepsilon > 0$ , on a  $\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) = \mathbb{P}(|X - \mathbb{E}[X]|^p \geq \varepsilon^p) \leq \frac{1}{\varepsilon^p} \mathbb{E}[|X - \mathbb{E}[X]|^p]$  en appliquant l'inégalité de Markov.  $\square$

*Exemple 1.3.17.* Si  $X \in L^2(\Omega)$  centrée et réduite,  $\mathbb{P}(|X| \geq 10) = \mathbb{P}(|X - \mathbb{E}[X]| \geq 10) \leq \frac{\text{Var}(X)}{100} = \frac{1}{100}$ .

**Proposition 1.3.18** (Inégalité de Cauchy–Schwarz). Soient  $X$  et  $Y \in L^2(\Omega)$  alors  $XY \in L^1(\Omega)$  et  $|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]}$ .

*Démonstration.* On a  $0 \leq |X||Y| \leq X^2 + Y^2$  donc  $\mathbb{E}[|X||Y|] \leq \mathbb{E}[X^2] + \mathbb{E}[Y^2] < +\infty$  et donc  $XY \in L^1(\Omega)$ . Ensuite, pour tout  $\lambda \in \mathbb{R}$ , on a  $X + \lambda Y \in L^2(\Omega)$  et

$$0 \leq \mathbb{E}[(X + \lambda Y)^2] = \mathbb{E}[X^2] + 2\lambda\mathbb{E}[XY] + \lambda^2\mathbb{E}[Y^2].$$

Ce polynôme de degré 2 en  $\lambda$  reste positif, donc son discriminant  $4\mathbb{E}[XY]^2 - 4\mathbb{E}[X^2]\mathbb{E}[Y^2]$  est négatif ou nul. Donc  $\mathbb{E}[XY]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$ .  $\square$

**Définition 1.3.19** (Covariance). Soient  $X$  et  $Y \in L^2(\Omega)$ , on appelle *covariance* de  $X$  et  $Y$  la quantité  $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ . Si  $\text{Cov}(X, Y) = 0$ , on dit que  $X$  et  $Y$  sont *décorrélées*.

**Proposition 1.3.20** (Propriété de la covariance). L'application  $(X, Y) \mapsto \text{Cov}(X, Y)$  est bilinéaire symétrique sur  $L^2(\Omega)$ . De plus, pour tout  $X$  et  $Y \in L^2(\Omega)$ , on a :

- $\text{Cov}(X, X) = \text{Var}(X) \geq 0$  ;
- $|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}$ , en particulier  $\text{Cov}(X, Y) = 0$  si  $X$  ou  $Y$  est constante p.s.;
- $\text{Var}(X + Y) = \text{Var}(X) + 2\text{Cov}(X, Y) + \text{Var}(Y)$ .

*Démonstration.* Exercice.  $\square$

## 1.4 Indépendance

De nouveau, on fixe un espace de probabilité abstrait  $(\Omega, \mathcal{A}, \mathbb{P})$  sur lequel seront définies toutes nos variables aléatoires.

### 1.4.1 Lois marginales et mesure produit

**Définition 1.4.1** (Marginales). Soit  $X = (X_1, \dots, X_n)$  un vecteur aléatoire à valeurs dans  $\mathbb{R}^n$ . Les v.a.r.  $(X_i)_{1 \leq i \leq n}$  sont appelées les *marginales* de  $X$ . Leurs lois  $(\mathbb{P}_{X_i})_{1 \leq i \leq n}$  sont appelées les *lois marginales* de la loi  $\mathbb{P}_X$  de  $X$ .

Soit  $i \in \llbracket 1, n \rrbracket$ , pour tout  $B \in \mathcal{B}(\mathbb{R})$ , on a

$$\mathbb{P}_{X_i}(B) = \mathbb{P}(X_i \in B) = \mathbb{P}(X \in \mathbb{R}^{i-1} \times B \times \mathbb{R}^{n-i}) = \mathbb{P}_X(\mathbb{R}^{i-1} \times B \times \mathbb{R}^{n-i}). \quad (1.4.1)$$

Donc, les lois marginales  $(\mathbb{P}_{X_i})_{1 \leq i \leq n}$  sont totalement déterminées par la loi  $\mathbb{P}_X$ . Dans le cas d'une loi discrète ou à densité, cela se traduit comme suit.

**Proposition 1.4.2.** Soit  $X = (X_1, \dots, X_n)$  un vecteur aléatoire à valeur dans  $\mathbb{R}^n$ . Soit  $i \in \llbracket 1, n \rrbracket$  et  $\Pi_i : (x_1, \dots, x_n) \mapsto x_i$ .

- Si  $X$  est discret à valeur dans  $D$  dénombrable alors  $X_i$  est discrète et, pour tout  $t \in \mathbb{R}$  on a  $\mathbb{P}(X_i = t) = \sum_{x \in \Pi_i^{-1}(\{t\})} \mathbb{P}(X = x)$ .
- Si  $X$  admet une densité  $f$ , alors  $X_i$  admet la densité

$$f_i : t \mapsto \int_{\mathbb{R}^{n-1}} f(x_1, \dots, x_{i-1}, t, x_{i+1}, \dots, x_n) dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_n.$$

*Démonstration.* Dans le cas discret,  $X_i = \Pi_i(X)$  est une v.a.r. discrète. Pour tout  $t \in \mathbb{R}$ , on a

$$\mathbb{P}(X_i = t) = \mathbb{P}(X \in \Pi_i^{-1}(\{t\})) = \sum_{x \in \Pi_i^{-1}(\{t\})} \mathbb{P}(X = x).$$

Dans le cas à densité, pour tout  $a \leq b$  on a d'après (1.4.1),

$$\begin{aligned} \mathbb{P}_{X_i}([a; b]) &= \mathbb{P}_X(\mathbb{R}^{i-1} \times [a; b] \times \mathbb{R}^{n-i}) = \int_{\mathbb{R}^{i-1} \times [a; b] \times \mathbb{R}^{n-i}} f(x_1, \dots, x_n) dx_1 \dots dx_n \\ &= \int_a^b \left( \int_{\mathbb{R}^{n-1}} f(x_1, \dots, x_n) dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_n \right) dx_i = \int_a^b f_i(x_i) dx_i, \end{aligned}$$

où on a utilisé le théorème de Fubini pour  $f$  positive. Le même calcul, en remplaçant  $[a; b]$  par  $\mathbb{R}$  montre que  $f_i$  est bien d'intégrale 1. Donc  $X_i$  admet  $f_i$  pour densité, voir définition 1.1.14.  $\square$

La loi d'un vecteur aléatoire n'est pas déterminée par ses lois marginales. Par exemple, considérons  $X = (X_1, X_2)$  et  $Y = (Y_1, Y_2)$  deux vecteurs aléatoires discrets à valeurs dans  $\{0; 1\}^2$ , où  $X$  est uniforme et la loi de  $Y$  est décrite par :

$$\mathbb{P}(Y = (0; 0)) = \frac{1}{2} = \mathbb{P}(Y = (1; 1)) \quad \text{et} \quad \mathbb{P}(Y = (1; 0)) = 0 = \mathbb{P}(Y = (0; 1)).$$

Alors  $X_1, X_2, Y_1, Y_2$  sont des v.a.r. discrètes à valeurs dans  $\{0; 1\}$  et on a

$$\begin{aligned} \mathbb{P}(X_1 = 0) &= \mathbb{P}(X = (0; 0)) + \mathbb{P}(X = (0; 1)) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}, & \mathbb{P}(X_1 = 1) &= 1 - \mathbb{P}(X_1 = 0) = \frac{1}{2}, \\ \mathbb{P}(Y_1 = 0) &= \mathbb{P}(Y = (0; 0)) + \mathbb{P}(Y = (0; 1)) = \frac{1}{2} + 0 = \frac{1}{2}, & \mathbb{P}(Y_1 = 1) &= 1 - \mathbb{P}(Y_1 = 0) = \frac{1}{2}. \end{aligned}$$

Donc  $X_1$  et  $Y_1$  sont des variables de Bernoulli de paramètre  $\frac{1}{2}$ . Le même raisonnement montre que  $X_2$  et  $Y_2 \sim \mathcal{B}(\frac{1}{2})$  aussi. Donc  $\mathbb{P}_X$  et  $\mathbb{P}_Y$  sont différentes mais ont les mêmes lois marginales.

**Théorème 1.4.3** (Carathéodory). *Soient  $(\mathbb{P}_i)_{1 \leq i \leq n}$  des probabilités sur  $\mathbb{R}$ , il existe une unique mesure de probabilité  $\mathbb{P}$  sur  $\mathbb{R}^n$  telle que, pour tout  $B_1, \dots, B_n \in \mathcal{B}(\mathbb{R})$ ,  $\mathbb{P}(\prod_{i=1}^n B_i) = \prod_{i=1}^n \mathbb{P}_i(B_i)$ .*

*Démonstration.* Admis. L'unicité se déduit du lemme de classe monotone, puisqu'on définit en particulier  $\mathbb{P}$  sur les pavés fermés. L'existence est difficile, voir [3, app. A].  $\square$

*Remarque 1.4.4.* Il est vrai, mais pas immédiat que si  $B_1, \dots, B_n \in \mathcal{B}(\mathbb{R})$  alors  $\prod_{i=1}^n B_i \in \mathcal{B}(\mathbb{R}^n)$ .

**Définition 1.4.5** (Mesure produit). La mesure de probabilité  $\mathbb{P}$  sur  $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$  est appelée la *mesure produit* des  $(\mathbb{P}_i)_{1 \leq i \leq n}$ , et elle est notée  $\mathbb{P}_1 \otimes \dots \otimes \mathbb{P}_n$  ou  $\bigotimes_{i=1}^n \mathbb{P}_i$ .

*Exemple 1.4.6.* Soient  $(\mathbb{P}_i)_{1 \leq i \leq n}$  des mesures de probabilité sur  $\mathbb{R}$  et  $\mathbb{P} = \bigotimes_{i=1}^n \mathbb{P}_i$ .

- Si pour tout  $i \in \llbracket 1, n \rrbracket$  la proba  $\mathbb{P}_i$  admet une densité  $f_i$ , alors pour tout  $a_1 \leq b_1; \dots; a_n \leq b_n$ ,

$$\mathbb{P} \left( \prod_{i=1}^n [a_i; b_i] \right) = \prod_{i=1}^n \mathbb{P}_i([a_i; b_i]) = \prod_{i=1}^n \int_{a_i}^{b_i} f_i(x_i) dx_i = \int_{\prod_{i=1}^n [a_i; b_i]} \prod_{i=1}^n f_i(x_i) dx_1 \dots dx_n.$$

Donc  $\mathbb{P}$  admet la densité  $\bigotimes_{i=1}^n f_i : (x_1, \dots, x_n) \mapsto \prod_{i=1}^n f_i(x_i)$ , voir définition 1.1.14.

- Si pour tout  $i \in \llbracket 1, n \rrbracket$  la proba  $\mathbb{P}_i$  est discrète, i.e. il existe  $D_i \subset \mathbb{R}^n$  dénombrable telle que  $\mathbb{P}_i(D_i) = 1$ . Alors  $\prod_{i=1}^n D_i$  est dénombrable et  $\mathbb{P}(\prod_{i=1}^n D_i) = \prod_{i=1}^n \mathbb{P}_i(D_i) = 1$ . Donc  $\mathbb{P}$  est discrète. De plus, pour tout  $(x_1, \dots, x_n) \in \prod_{i=1}^n D_i$  on a  $\mathbb{P}(\{(x_1, \dots, x_n)\}) = \prod_{i=1}^n \mathbb{P}_i(\{x_i\})$ .

**Lemme 1.4.7.** Soient  $(\mathbb{P}_i)_{1 \leq i \leq n}$  des mesures de probabilité sur  $\mathbb{R}$  et  $\mathbb{P} = \bigotimes_{i=1}^n \mathbb{P}_i$ . Alors les  $(\mathbb{P}_i)_{1 \leq i \leq n}$  sont les marginales de  $\mathbb{P}$ .

*Démonstration.* Soit  $X = (X_1, \dots, X_n)$  un vecteur aléatoire de loi  $\mathbb{P}$ . Soit  $i \in \llbracket 1, n \rrbracket$ , on doit vérifier que  $\mathbb{P}_{X_i} = \mathbb{P}_i$ . Or, d'après (1.4.1), pour tout  $B \in \mathcal{B}(\mathbb{R}^n)$ ,

$$\mathbb{P}_{X_i}(B) = \mathbb{P}_X(\mathbb{R}^{i-1} \times B \times \mathbb{R}^{n-i}) = \mathbb{P}(\mathbb{R}^{i-1} \times B \times \mathbb{R}^{n-i}) = \mathbb{P}_i(B) \times \prod_{j \neq i} \mathbb{P}_j(\mathbb{R}) = \mathbb{P}_i(B). \quad \square$$

## 1.4.2 Évènements et variables indépendantes

**Définition 1.4.8** (Évènements indépendants). Soient  $(A_i)_{i \in I}$  une famille d'éléments de  $\mathcal{A}$ . On dit que ces évènements sont *indépendants* si, pour tout  $J \subset I$  fini,  $\mathbb{P}(\bigcap_{j \in J} A_j) = \prod_{j \in J} \mathbb{P}(A_j)$ .

**Définition 1.4.9** (Variables aléatoires indépendantes). Soient  $(X_i)_{i \in I}$  une famille de v.a. définies sur un même espace de probabilité  $(\Omega, \mathcal{A}, \mathbb{P})$  à valeurs dans des espaces mesurés  $(E_i, \mathcal{B}_i)_{i \in I}$ . On dit que les v.a.  $(X_i)_{i \in I}$  sont *indépendantes* si, pour tout  $J \subset I$  fini, pour tout  $(B_j)_{j \in J} \in \prod_{j \in J} \mathcal{B}_j$  on a

$$\mathbb{P}(\forall j \in J, X_j \in B_j) = \prod_{j \in J} \mathbb{P}(X_j \in B_j).$$

En particulier, les v.a.r.  $X_1, \dots, X_n$  sont indépendantes si, pour tout  $B_1, \dots, B_n \in \mathcal{B}(\mathbb{R})$  on a  $\mathbb{P}(\forall i \in \llbracket 1, n \rrbracket, X_i \in B_i) = \prod_{i=1}^n \mathbb{P}(X_i \in B_i)$ .

Vérifier l'indépendance d'une famille de v.a. c'est vérifier l'indépendance de toutes les sous-familles finis. Dans le cas de v.a.r., on a plusieurs critères pour prouver cette indépendance.

**Proposition 1.4.10.** Soient  $X_1, \dots, X_n$  des v.a.r. et  $X = (X_1, \dots, X_n)$ . Les propositions suivantes sont équivalentes.

1.  $X_1, \dots, X_n$  sont indépendantes.
2. La loi  $\mathbb{P}_X$  de  $X$  est la loi produit  $\bigotimes_{i=1}^n \mathbb{P}_{X_i}$  des lois des  $(X_i)_{1 \leq i \leq n}$ .
3. Pour tout  $a_1 \leq b_1, \dots, a_n \leq b_n$  on a  $\mathbb{P}(\forall i \in \llbracket 1, n \rrbracket, X_i \in [a_i; b_i]) = \prod_{i=1}^n \mathbb{P}(X_i \in [a_i; b_i])$ .
4. Pour toutes fonctions  $f_1, \dots, f_n$  de  $\mathbb{R}$  dans  $[0; +\infty[$  on a  $\mathbb{E}[\prod_{i=1}^n f_i(X_i)] = \prod_{i=1}^n \mathbb{E}[f_i(X_i)]$ .
5. Pour toutes fonctions  $f_1, \dots, f_n$  de  $\mathbb{R}$  dans  $\mathbb{R}$  telles que les  $(f_i(X_i))_{1 \leq i \leq n}$  soient  $L^1$  on a  $\prod_{i=1}^n f_i(X_i)$  intégrable et  $\mathbb{E}[\prod_{i=1}^n f_i(X_i)] = \prod_{i=1}^n \mathbb{E}[f_i(X_i)]$ .
6. En notant  $F_X$  (resp.  $F_{X_i}$ ) la fonction de répartition de  $X$  (resp.  $X_i$ ), on a  $F_X = \bigotimes_{i=1}^n F_{X_i}$ .

*Démonstration.* Si (1) est vraie alors, en appliquant la définition 1.4.9 avec  $B_i = [a_i; b_i]$  pour tout  $i \in \llbracket 1, n \rrbracket$ , on obtient (3). Si (3) est vraie, alors pour tout pavé fermé  $P = \prod_{i=1}^n [a_i; b_i]$  on a

$$\mathbb{P}_X(P) = \mathbb{P}(\forall i \in \llbracket 1, n \rrbracket, X_i \in [a_i; b_i]) = \prod_{i=1}^n \mathbb{P}(X_i \in [a_i; b_i]) = \prod_{i=1}^n \mathbb{P}_{X_i}([a_i; b_i]) = \left( \bigotimes_{i=1}^n \mathbb{P}_{X_i} \right)(P).$$

Donc  $\mathbb{P}_X$  et  $\bigotimes_{i=1}^n \mathbb{P}_{X_i}$  coïncident sur les pavés fermés. Comme la famille des pavés fermés engendre  $\mathcal{B}(\mathbb{R}^n)$  et est stable par intersection finie, ces mesures sont égales par le lemme de classe monotone, ce qui prouve (2). Enfin, si (2) est vraie alors, pour tout  $B_1, \dots, B_n \in \mathcal{B}(\mathbb{R})$  on a :

$$\mathbb{P}(\forall i \in \llbracket 1, n \rrbracket, X_i \in B_i) = \mathbb{P}\left(X \in \prod_{i=1}^n B_i\right) = \mathbb{P}_X\left(\prod_{i=1}^n B_i\right) = \prod_{i=1}^n \mathbb{P}_{X_i}(B_i) = \prod_{i=1}^n \mathbb{P}(X_i \in B_i).$$

Ainsi, (1) implique (3) implique (2) implique (1), ce qui prouve l'équivalence entre (1), (2) et (3).

On admet l'équivalence avec (4) et (5) qui vient de la théorie de la mesure. Si on suppose (2), en appliquant la définition 1.4.9 avec  $B_i = ] - \infty; x_i]$ , il vient que, pour tout  $x_1, \dots, x_n \in \mathbb{R}$  on a

$$F_X(x_1, \dots, x_n) = \mathbb{P}_X \left( \prod_{i=1}^n ] - \infty; x_i] \right) = \prod_{i=1}^n \mathbb{P}_{X_i}(] - \infty; x_i]) = \prod_{i=1}^n F_{X_i}(x_i),$$

ce qui prouve (6). Inversement, notons  $Y$  un vecteur aléatoire de loi  $\bigotimes_{i=1}^n \mathbb{P}_{X_i}$ . Si on suppose (6) alors, en appliquant (2) implique (6) à  $Y$ , la fonction de répartition  $F_Y$  de  $Y$  est  $F_Y = \bigotimes_{i=1}^n F_{X_i} = F_X$ . Donc  $\mathbb{P}_X = \mathbb{P}_Y = \bigotimes_{i=1}^n \mathbb{P}_{X_i}$  d'après le lemme 1.2.16.  $\square$

Une application de la dernière caractérisation est de déterminer la loi du maximum d'une famille de variables aléatoires. Soient  $X_1, \dots, X_n$  des v.a.r. indépendantes et de même loi, dont on note  $F$  la fonction de répartition. Soit  $Y = \max_{1 \leq i \leq n} X_i$  et  $F_Y$  sa fonction de répartition. Pour tout  $x \in \mathbb{R}$ ,

$$F_Y(x) = \mathbb{P} \left( \max_{1 \leq i \leq n} X_i \leq x \right) = \mathbb{P}(\forall i \in \llbracket 1, n \rrbracket, X_i \leq x) = F_X(x, \dots, x) = \prod_{i=1}^n F_{X_i}(x) = F(x)^n.$$

Donc  $F_Y = F^n$ , ce qui décrit totalement la loi de  $Y$  en fonction de la loi commune des  $(X_i)_{1 \leq i \leq n}$ .

**Lemme 1.4.11.** *Soient  $X_1, \dots, X_n$  des v.a.r.  $L^2$  indépendantes. Pour tout  $i, j \in \llbracket 1, n \rrbracket$  tels que  $i \neq j$ , on a  $\text{Cov}(X_i, X_j) = 0$ . De plus  $\text{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \text{Var}(X_i)$ .*

*Démonstration.* Quitte à remplacer  $X_i$  par  $X_i - \mathbb{E}[X_i]$  pour tout  $i \in \llbracket 1, n \rrbracket$ , ce qui ne modifie pas les variances et covariances considérées, on peut supposer les  $(X_i)_{1 \leq i \leq n}$  centrées. Soit  $i \neq j$  dans  $\llbracket 1, n \rrbracket$ , on a  $\text{Cov}(X_i, X_j) = \mathbb{E}[X_i X_j] = \mathbb{E}[X_i] \mathbb{E}[X_j] = 0$ , en utilisant l'indépendance de  $X_i$  et  $X_j$  sous la forme (5), voir proposition 1.4.10. Ensuite,  $(\sum_{i=1}^n X_i)^2 = \sum_{1 \leq i, j \leq n} X_i X_j$ , donc

$$\text{Var} \left( \sum_{i=1}^n X_i \right) = \mathbb{E} \left[ \left( \sum_{i=1}^n X_i \right)^2 \right] = \sum_{1 \leq i, j \leq n} \mathbb{E}[X_i X_j] = \sum_{1 \leq i, j \leq n} \text{Cov}(X_i, X_j) = \sum_{i=1}^n \text{Var}(X_i),$$

en utilisant le premier point pour annuler les  $\text{Cov}(X_i, X_j)$  avec  $i \neq j$ .  $\square$

Soient  $\mathbb{P}_1, \dots, \mathbb{P}_n$  des mesures de probabilités sur  $\mathbb{R}$ . Par le théorème 1.4.3, la mesure produit  $\mathbb{P} = \bigotimes_{i=1}^n \mathbb{P}_i$  existe. Soit  $X = (X_1, \dots, X_n)$  un vecteur aléatoire de loi  $\mathbb{P}$ . Pour tout  $i \in \llbracket 1, n \rrbracket$ ,  $X_i$  est une v.a.r. de loi  $\mathbb{P}_i$  d'après le lemme 1.4.7. De plus les  $(X_i)_{1 \leq i \leq n}$  sont indépendantes. Donc, il existe une famille  $(X_i)_{1 \leq i \leq n}$  indépendantes et de lois prescrites. Ceci reste vrai si on considère des suites de variables aléatoires.

**Théorème 1.4.12** (Kolmogorov). *Soit  $(\mathbb{P}_k)_{k \in \mathbb{N}}$  une suite de mesures de probabilités sur  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ . Il existe un espace de probabilités  $(\Omega, \mathcal{A}, \mathbb{P})$  est une suite  $(X_k)_{k \in \mathbb{N}}$  de v.a.r. sur  $(\Omega, \mathcal{A}, \mathbb{P})$  telle que :*

- pour tout  $k \in \mathbb{N}$ ,  $\mathbb{P}_{X_k} = \mathbb{P}_k$  ;
- les  $(X_k)_{k \in \mathbb{N}}$  sont indépendantes.

*Démonstration.* Admis, voir [3, app. A].  $\square$

**Définition 1.4.13** (Suites de v.a.i.i.d.). Si  $(X_k)_{k \in \mathbb{N}}$  est une suite de variables aléatoires indépendantes et de même loi, et on dit que c'est une suite de *variables aléatoires indépendantes et identiquement distribuées* (v.a.i.i.d.).

## Chapitre 2

# Convergence de suites de variables aléatoires

Dans ce chapitre, toutes les variables aléatoires définies sur un espace de probabilité  $(\Omega, \mathcal{A}, \mathbb{P})$ . On note  $\|\cdot\|$  la norme euclidienne usuelle de  $\mathbb{R}^n$ .

### 2.1 Convergence en probabilité

**Définition 2.1.1** (Convergence en probabilité). Soient  $X$  et  $(X_k)_{k \in \mathbb{N}}$  des v.a. à valeurs dans  $\mathbb{R}^n$ . On dit que  $(X_k)_{k \in \mathbb{N}}$  converge en probabilité vers  $X$ , et on note  $X_k \xrightarrow[k \rightarrow +\infty]{\mathbb{P}} X$ , si pour tout  $\varepsilon > 0$  on a  $\mathbb{P}(\|X_k - X\| > \varepsilon) \xrightarrow[k \rightarrow +\infty]{} 0$ .

*Exemple 2.1.2.* Supposons que, pour tout  $k \in \mathbb{N}^*$ ,  $X_k \sim \mathcal{U}([0; \frac{1}{k}])$ . Soit  $\varepsilon > 0$ , pour tout  $k > \frac{1}{\varepsilon}$ ,  $[0; \frac{1}{k}] \subset [-\varepsilon; \varepsilon]$  et donc  $\mathbb{P}(|X_k| > \varepsilon) = 0$ . Donc  $\mathbb{P}(|X_k| > \varepsilon) \xrightarrow[k \rightarrow +\infty]{} 0$  pour tout  $\varepsilon > 0$ , et  $X_k \xrightarrow[k \rightarrow +\infty]{\mathbb{P}} 0$ .

**Proposition 2.1.3.** Soient  $(X_k)_{k \in \mathbb{N}}$  une suite de v.a. à valeurs dans  $\mathbb{R}^n$  qui converge en probabilité vers  $X$  et  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  une fonction continue, alors  $f(X_k) \xrightarrow[k \rightarrow +\infty]{\mathbb{P}} f(X)$  dans  $\mathbb{R}^m$ .

*Démonstration.* Soit  $\varepsilon > 0$ , on veut montrer que  $\mathbb{P}(\|f(X_k) - f(X)\| > \varepsilon) \xrightarrow[k \rightarrow +\infty]{} 0$ .

Soit  $\delta > 0$ , comme  $\mathbb{P}_X(B(0, R)^c) \xrightarrow[R \rightarrow +\infty]{} \mathbb{P}_X(\emptyset) = 0$ , il existe  $R > 0$  tel que  $\mathbb{P}_X(B(0, R)^c) \leq \delta$ .

Sur la boule compacte  $\overline{B(0, 2R)}$  la fonction  $f$  est uniformément continue. Il existe donc  $\eta > 0$  tel que si  $x, y \in \overline{B(0, 2R)}$  et  $\|x - y\| \leq \eta$  alors  $\|f(x) - f(y)\| \leq \varepsilon$ . On peut supposer  $\eta \leq R$ .

Si  $\|x\| < R$  et  $\|y - x\| \leq \eta$ , alors  $x, y \in \overline{B(0, 2R)}$  et  $\|f(x) - f(y)\| \leq \varepsilon$ . Par contraposée, si  $\|f(x) - f(y)\| > \varepsilon$  alors  $x \in B(0, R)^c$  ou  $\|y - x\| > \eta$ . Donc, pour tout  $k$  assez grand,

$$\mathbb{P}(\|f(X_k) - f(X)\| > \varepsilon) \leq \mathbb{P}(X \in B(0, R)^c) + \mathbb{P}(\|X_k - X\| > \eta) \leq \delta + \mathbb{P}(\|X_k - X\| > \eta) \leq 2\delta. \quad \square$$

*Exemple 2.1.4.* Soient  $(X_k)$  et  $(Y_k)$  deux suites de v.a.r. telles que  $X_k \xrightarrow[k \rightarrow +\infty]{\mathbb{P}} X$  et  $Y_k \xrightarrow[k \rightarrow +\infty]{\mathbb{P}} Y$ . Soit

$\varepsilon > 0$ . Si  $|X_k - X| \leq \frac{\varepsilon}{2}$  et  $|Y_k - Y| \leq \frac{\varepsilon}{2}$  alors  $\|(X_k, Y_k) - (X, Y)\| = \left(|X_k - X|^2 + |Y_k - Y|^2\right)^{\frac{1}{2}} \leq \varepsilon$ . Donc  $\|(X_k, Y_k) - (X, Y)\| > \varepsilon$  implique  $|X_k - X| > \frac{\varepsilon}{2}$  ou  $|Y_k - Y| > \frac{\varepsilon}{2}$ , donc

$$\mathbb{P}(\|(X_k, Y_k) - (X, Y)\| > \varepsilon) \leq \mathbb{P}\left(|X_k - X| > \frac{\varepsilon}{2}\right) + \mathbb{P}\left(|Y_k - Y| > \frac{\varepsilon}{2}\right) \xrightarrow[k \rightarrow +\infty]{} 0.$$

Donc  $(X_k, Y_k) \xrightarrow[k \rightarrow +\infty]{\mathbb{P}} (X, Y)$  et, par continuité de  $+$ , on a  $X_k + Y_k \xrightarrow[k \rightarrow +\infty]{\mathbb{P}} X + Y$ .

**Théorème 2.1.5** (Loi faible des grands nombres). *Soit  $(X_k)_{k \geq 1}$  une suite de v.a.i.i.d.  $L^1$  réelles d'espérance  $\mu \in \mathbb{R}$ . Pour tout  $n \in \mathbb{N}^*$ , on note  $S_n = \frac{1}{n} \sum_{k=1}^n X_k$ . Alors  $S_n \xrightarrow[n \rightarrow +\infty]{\mathbb{P}} \mu$ .*

*Démonstration.* On fait la preuve sous l'hypothèse supplémentaire que les  $(X_k)_{k \geq 1}$  sont  $L^2$ .

Soit  $n \in \mathbb{N}^*$ , alors la v.a.r.  $S_n$  est  $L^2$  comme combinaison linéaire de v.a.r.  $L^2$ . Par linéarité de l'espérance,  $\mathbb{E}[S_n] = \frac{1}{n} \sum_{k=1}^n \mathbb{E}[X_k] = \mu$ . Enfin,  $X_1, \dots, X_n$  étant indépendantes,

$$\text{Var}(S_n) = \text{Var}\left(\frac{1}{n} \sum_{k=1}^n X_k\right) = \frac{1}{n^2} \sum_{k=1}^n \text{Var}(X_k) = \frac{\sigma^2}{n},$$

où  $\sigma^2$  est la variance commune des  $(X_k)_{k \geq 1}$ . Soit  $\varepsilon > 0$ , par l'inégalité de Chebyshev,

$$\mathbb{P}(|S_n - \mu| > \varepsilon) = \mathbb{P}(|S_n - \mathbb{E}[S_n]| > \varepsilon) \leq \frac{1}{\varepsilon^2} \text{Var}(S_n) = \frac{\sigma^2}{n\varepsilon^2} \xrightarrow[n \rightarrow +\infty]{} 0. \quad \square$$

## 2.2 Convergence presque sûre

**Définition 2.2.1** (Convergence presque sûre). Soient  $X$  et  $(X_k)_{k \in \mathbb{N}}$  des v.a. à valeurs dans  $\mathbb{R}^n$ . On dit que  $(X_k)_{k \in \mathbb{N}}$  converge presque sûrement vers  $X$ , et on note  $X_k \xrightarrow[k \rightarrow +\infty]{p.s.} X$ , si

$$\mathbb{P}\left(\left\{\omega \in \Omega \mid X_k(\omega) \xrightarrow[k \rightarrow +\infty]{} X(\omega)\right\}\right) = \mathbb{P}\left(X_k \xrightarrow[k \rightarrow +\infty]{} X\right) = 1.$$

En tant que fonctions sur  $\Omega$ , c'est la convergence simple de  $(X_k)_{k \in \mathbb{N}}$  vers  $X$  en dehors d'un évènement de probabilité nulle.

**Lemme 2.2.2.** *Si  $X_k \xrightarrow[k \rightarrow +\infty]{p.s.} X$  dans  $\mathbb{R}^n$  et  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  est continue alors  $f(X_k) \xrightarrow[k \rightarrow +\infty]{p.s.} f(X)$ . En particulier, si  $X_k \xrightarrow[k \rightarrow +\infty]{p.s.} X$  et  $Y_k \xrightarrow[k \rightarrow +\infty]{p.s.} Y$ , alors  $X_k + Y_k \xrightarrow[k \rightarrow +\infty]{p.s.} X + Y$ .*

*Démonstration.* Exercice. □

**Lemme 2.2.3.** *Si  $X_k \xrightarrow[k \rightarrow +\infty]{p.s.} X$  dans  $\mathbb{R}^n$  alors  $X_k \xrightarrow[k \rightarrow +\infty]{\mathbb{P}} X$ .*

*Démonstration.* Soit  $\varepsilon > 0$ . Soit  $\omega \in \Omega$  tel que  $X_k(\omega) \xrightarrow[k \rightarrow +\infty]{} X(\omega)$ , il existe  $m \in \mathbb{N}$  tel que pour tout  $k \geq m$ ,  $\|X_k(\omega) - X(\omega)\| \leq \varepsilon$ . Donc  $\left\{X_k \xrightarrow[k \rightarrow +\infty]{} X\right\} \subset \bigcup_{m \in \mathbb{N}} \{\forall k \geq m, \|X_k - X\| \leq \varepsilon\}$  et

$$1 = \mathbb{P}\left(X_k \xrightarrow[k \rightarrow +\infty]{} X\right) \leq \mathbb{P}\left(\bigcup_{m \in \mathbb{N}} \{\forall k \geq m, \|X_k - X\| \leq \varepsilon\}\right) = \lim_{m \rightarrow +\infty} \mathbb{P}(\forall k \geq m, \|X_k - X\| \leq \varepsilon).$$

Pour tout  $m \in \mathbb{N}$ , on a  $\mathbb{P}(\forall k \geq m, \|X_k - X\| \leq \varepsilon) \leq \mathbb{P}(\|X_m - X\| \leq \varepsilon) \leq 1$ . Donc, par le théorème des gendarmes,  $\mathbb{P}(\|X_m - X\| \leq \varepsilon) \xrightarrow[m \rightarrow +\infty]{} 1$  et  $\mathbb{P}(\|X_m - X\| > \varepsilon) \xrightarrow[m \rightarrow +\infty]{} 0$ . □

En général, la réciproque du lemme 2.2.3 est fautive. Cependant, on peut parfois obtenir une convergence p.s. si on a une convergence en proba quantitative assez rapide, grâce au lemme suivant.

**Lemme 2.2.4** (Borel–Cantelli). Soit  $(A_k)_{k \in \mathbb{N}}$  des évènements de  $\mathcal{A}$ . Si  $\sum_{k \in \mathbb{N}} \mathbb{P}(A_k) < +\infty$  alors  $\mathbb{P}(\bigcap_{n \in \mathbb{N}} \bigcup_{k \geq n} A_k) = 0$ .

*Démonstration.* Pour tout  $p \in \mathbb{N}$ , on a  $\mathbb{P}\left(\bigcap_{n \in \mathbb{N}} \bigcup_{k \geq n} A_k\right) \leq \mathbb{P}\left(\bigcup_{k \geq p} A_k\right) \leq \sum_{k \geq p} \mathbb{P}(A_k) \xrightarrow{p \rightarrow +\infty} 0$ .  $\square$

**Théorème 2.2.5** (Loi forte des grands nombres). Soit  $(X_k)_{k \geq 1}$  une suite de v.a.i.i.d.  $L^1$  réelles d'espérance  $\mu \in \mathbb{R}$ . Pour tout  $n \in \mathbb{N}^*$ , on note  $S_n = \frac{1}{n} \sum_{k=1}^n X_k$ . Alors  $S_n \xrightarrow[n \rightarrow +\infty]{\text{p.s.}} \mu$ .

*Remarque 2.2.6.* D'après le lemme 2.2.3, la loi forte (thm. 2.2.5) implique la loi faible (thm. 2.1.5).

*Démonstration sous hypothèse  $L^4$ .* Il suffit de prouver le résultat lorsque  $\mu = 0$ . En effet, dans le cas général, si on pose  $\tilde{X}_k = X_k - \mu$  pour tout  $k \geq 1$  alors les  $(\tilde{X}_k)_{k \geq 1}$  sont des v.a.i.i.d.  $L^1$  d'espérance nulle. Pour tout  $n \in \mathbb{N}^*$ , on a  $\tilde{S}_n = \frac{1}{n} \sum_{k=1}^n \tilde{X}_k = \frac{1}{n} \sum_{k=1}^n (X_k - \mu) = S_n - \mu$ . Si le résultat est établi dans le cas  $\mu = 0$ , alors  $\tilde{S}_n \xrightarrow[n \rightarrow +\infty]{\text{p.s.}} 0$ , ce qui implique  $S_n \xrightarrow[n \rightarrow +\infty]{\text{p.s.}} \mu$ .

Reste à prouver le résultat lorsque  $\mu = 0$ . On le fait sous l'hypothèse supplémentaire que les  $(X_k)_{k \geq 1}$  sont  $L^4$ . On note  $\sigma^2 = \mathbb{E}[X_1^2]$  et  $\kappa = \mathbb{E}[X_1^4] \geq 0$  leur moment (centré) d'ordre 4. Pour tout  $n \in \mathbb{N}^*$ , la v.a.r.  $S_n$  est  $L^4$  comme combinaison linéaire de variables  $L^4$  et, par linéarité de l'espérance,  $\mathbb{E}[S_n] = \frac{1}{n} \sum_{k=1}^n \mathbb{E}[X_k] = 0$ . On montrera plus loin que,

$$\mathbb{E}\left[\left(\sum_{k=1}^n X_k\right)^4\right] = 3n(n-1)\sigma^4 + n\kappa, \quad (2.2.1)$$

ce qui implique que  $\mathbb{E}[S_n^4] = \frac{3n(n-1)\sigma^4 + n\kappa}{n^4} \leq \frac{3\sigma^4 + \kappa}{n^2}$ . En admettant ceci pour l'instant, l'inégalité de concentration 1.3.16 pour le moment d'ordre 4 donne :

$$\mathbb{P}\left(|S_n| > n^{-\frac{1}{8}}\right) = \mathbb{P}\left(|S_n - \mathbb{E}[S_n]| > n^{-\frac{1}{8}}\right) \leq n^{\frac{1}{2}} \mathbb{E}[|S_n - \mathbb{E}[S_n]|^4] = n^{\frac{1}{2}} \mathbb{E}[S_n^4] \leq \frac{3\sigma^4 + \kappa}{n^{\frac{3}{2}}}.$$

Cette quantité étant sommable, on a  $\mathbb{P}\left(\bigcap_{N \geq 1} \bigcup_{n \geq N} \left\{|S_n| > n^{-\frac{1}{8}}\right\}\right) = 0$  par Borel–Cantelli 2.2.4. En passant au complémentaire,  $\mathbb{P}\left(\bigcup_{N \geq 1} \bigcap_{n \geq N} \left\{|S_n| \leq n^{-\frac{1}{8}}\right\}\right) = 1$ . Donc, presque sûrement, il existe  $N \geq 1$  tel que, pour tout  $n \geq N$ ,  $|S_n| \leq n^{-\frac{1}{8}}$ . En particulier,  $S_n \xrightarrow[n \rightarrow +\infty]{\text{p.s.}} 0$ .

Il reste à prouver (2.2.1), ce qu'on fait par récurrence sur  $n$ . Pour  $n = 1$ , on a bien  $\mathbb{E}[X_1^4] = \kappa$ . Supposons le résultat vrai pour  $n \geq 1$ . On a

$$\left(\sum_{k=1}^{n+1} X_k\right)^4 = \left(\sum_{k=1}^n X_k\right)^4 + 4\left(\sum_{k=1}^n X_k\right)^3 X_{n+1} + 6\left(\sum_{k=1}^n X_k\right)^2 X_{n+1}^2 + 4\left(\sum_{k=1}^n X_k\right) X_{n+1}^3 + X_{n+1}^4.$$

On calcule l'espérance de chacun des termes. On a  $\mathbb{E}\left[\left(\sum_{k=1}^n X_k\right)^4\right] = 3n(n-1)\sigma^4 + n\kappa$  par hypothèse de récurrence. Par indépendance,  $\mathbb{E}\left[\left(\sum_{k=1}^n X_k\right)^3 X_{n+1}\right] = \mathbb{E}\left[\left(\sum_{k=1}^n X_k\right)^3\right] \mathbb{E}[X_{n+1}] = 0$ . De même,

$$\mathbb{E}\left[\left(\sum_{k=1}^n X_k\right)^2 X_{n+1}^2\right] = \text{Var}\left(\sum_{k=1}^n X_k\right) \text{Var}(X_{n+1}) = \left(\sum_{k=1}^n \text{Var}(X_k)\right) \text{Var}(X_{n+1}) = n\sigma^4.$$

Puis  $\mathbb{E}[(\sum_{k=1}^n X_k)X_{n+1}^3] = \mathbb{E}[\sum_{k=1}^n X_k]\mathbb{E}[X_{n+1}^3] = (\sum_{k=1}^n \mathbb{E}[X_k])\mathbb{E}[X_{n+1}^3] = 0$ . Et  $\mathbb{E}[X_{n+1}^4] = \kappa$  par définition. Ainsi,

$$\mathbb{E}\left[\left(\sum_{k=1}^{n+1} X_k\right)^4\right] = 3n(n-1)\sigma^4 + n\kappa + 6n\sigma^4 + \kappa = 3n\sigma^4(n-1+2) + (n+1)\kappa = 3(n+1)n\sigma^4 + (n+1)\kappa.$$

Cela prouve l'hérédité, conclut la récurrence, et termine la preuve. Pour une preuve sous hypothèse  $L^1$  et diverses variantes, voir [3, chap. 10.2].  $\square$

## 2.3 Convergence en loi

### 2.3.1 Définition et caractérisations

**Définition 2.3.1** (Convergence en loi). Soit  $\mathcal{C}_b^0(\mathbb{R}^n)$  l'espace des fonctions continues bornées de  $\mathbb{R}^n$  dans  $\mathbb{R}$ . Soient  $X$  et  $(X_k)_{k \in \mathbb{N}}$  des variables aléatoires à valeurs dans  $\mathbb{R}^n$ , on dit que  $(X_k)_{k \in \mathbb{N}}$  converge en loi vers  $X$ , et on note  $X_k \xrightarrow[k \rightarrow +\infty]{\text{loi}} X$  si, pour tout  $f \in \mathcal{C}_b^0(\mathbb{R}^n)$ , on a  $\mathbb{E}[f(X_k)] \xrightarrow[k \rightarrow +\infty]{} \mathbb{E}[f(X)]$ .

*Remarque 2.3.2.* Cette notion de convergence dépend uniquement des lois  $\mathbb{P}_X$  et  $(\mathbb{P}_{X_k})_{k \in \mathbb{N}}$ . En particulier, la v.a. limite n'est pas unique, seule sa loi l'est.

**Lemme 2.3.3.** Si  $X_k \xrightarrow[k \rightarrow +\infty]{\mathbb{P}} X$  (en particulier si  $X_k \xrightarrow[k \rightarrow +\infty]{p.s.} X$ ) alors  $X_k \xrightarrow[k \rightarrow +\infty]{\text{loi}} X$ .

*Démonstration.* Soient  $f \in \mathcal{C}_b^0(\mathbb{R}^n)$ . Soit  $\varepsilon > 0$ , on note  $A_k = \{\omega \in \Omega \mid |f(X_k(\omega)) - f(X(\omega))| > \varepsilon\}$  pour tout  $k \in \mathbb{N}$ . On a,  $\forall \omega \in \Omega$ ,  $|f(X_k(\omega)) - f(X(\omega))| \leq \varepsilon + 2\|f\|_\infty \mathbf{1}_{A_k}(\omega)$ , donc

$$\begin{aligned} |\mathbb{E}[f(X_k)] - \mathbb{E}[f(X)]| &= |\mathbb{E}[f(X_k) - f(X)]| \leq \mathbb{E}[|f(X_k) - f(X)|] \leq \varepsilon + 2\|f\|_\infty \mathbb{E}[\mathbf{1}_{A_k}] \\ &\leq \varepsilon + 2\|f\|_\infty \mathbb{P}(A_k) = \varepsilon + 2\|f\|_\infty \mathbb{P}(|f(X_k) - f(X)| > \varepsilon). \end{aligned}$$

Comme  $f$  est continue, on a  $f(X_k) \xrightarrow[k \rightarrow +\infty]{\mathbb{P}} f(X)$ . Donc  $2\|f\|_\infty \mathbb{P}(|f(X_k) - f(X)| > \varepsilon) \xrightarrow[k \rightarrow +\infty]{} 0$  et donc, pour tout  $k$  assez grand, cette quantité est inférieure à  $\varepsilon$  et  $|\mathbb{E}[f(X_k)] - \mathbb{E}[f(X)]| \leq 2\varepsilon$ . Donc  $\mathbb{E}[f(X_k)] \xrightarrow[k \rightarrow +\infty]{} \mathbb{E}[f(X)]$ . C'est vrai pour tout  $f \in \mathcal{C}_b^0(\mathbb{R}^n)$ , donc  $X_k \xrightarrow[k \rightarrow +\infty]{\text{loi}} X$ .  $\square$

**Théorème 2.3.4** (Portemanteau). Soient  $X$  et  $(X_k)_{k \in \mathbb{N}}$  des v.a. à valeurs dans  $\mathbb{R}^n$ ,  $X_k \xrightarrow[k \rightarrow +\infty]{\text{loi}} X$  si et seulement si, pour tout  $B \in \mathcal{B}(\mathbb{R}^n)$  tel que  $\mathbb{P}(X \in \partial B) = 0$ , on a  $\mathbb{P}(X_k \in B) \xrightarrow[k \rightarrow +\infty]{} \mathbb{P}(X \in B)$ .

*Démonstration.* Admis. Voir [3, chap. 14.1].  $\square$

*Exemple 2.3.5.* Si  $n = 1$  et  $X$  admet la densité  $f$ , pour tout  $a < b$  on a  $\partial[a; b] = [a; b] \setminus ]a; b[ = \{a; b\}$  et  $\mathbb{P}(X \in \{a; b\}) = 0$ . Donc si  $X_k \xrightarrow[k \rightarrow +\infty]{\text{loi}} X$  alors  $\mathbb{P}(X_k \in [a; b]) \xrightarrow[k \rightarrow +\infty]{} \int_a^b f(x) dx$  pour tout  $a < b$ .

**Corollaire 2.3.6.** Soient  $X$  et  $(X_k)_{k \in \mathbb{N}}$  des v.a. discrètes à valeurs dans  $D \subset \mathbb{R}^n$  dénombrable, alors  $X_k \xrightarrow[k \rightarrow +\infty]{\text{loi}} X$  si et seulement si, pour tout  $x \in D$ , on a  $\mathbb{P}(X_k = x) \xrightarrow[k \rightarrow +\infty]{} \mathbb{P}(X = x)$ .

*Démonstration.* Admis.  $\square$

**Proposition 2.3.7.** Soient  $X$  et  $(X_k)_{k \in \mathbb{N}}$  des v.a.r.,  $F$  et  $(F_k)_{k \in \mathbb{N}}$  leurs fonctions de répartition. Soit  $C = \{x \in \mathbb{R} \mid F \text{ est continue en } x\}$ . Alors  $X_k \xrightarrow[k \rightarrow +\infty]{\text{loi}} X \iff \forall x \in C, F_k(x) \xrightarrow[k \rightarrow +\infty]{} F(x)$ . Si de plus  $F$  est continue (i.e.  $C = \mathbb{R}$ ), alors ces conditions sont équivalentes à  $\|F_k - F\|_\infty \xrightarrow[k \rightarrow +\infty]{} 0$ .

*Démonstration.* Admis. Le dernier point se déduit d'un théorème de Dini.  $\square$

### 2.3.2 Fonction caractéristique

Dans la suite, on note  $\langle \cdot, \cdot \rangle$  le produit scalaire canonique de  $\mathbb{R}^n$ .

**Définition 2.3.8** (Fonction caractéristique). Soit  $X = (X_1, \dots, X_n)$  une variable aléatoire à valeur dans  $\mathbb{R}^n$ , pour tout  $t = (t_1, \dots, t_n) \in \mathbb{R}^n$  la v.a. complexe  $e^{i\langle t, X \rangle} = e^{i\sum_{j=1}^n t_j X_j}$  est bornée donc  $L^1$ . La fonction caractéristique de  $X$  est la fonction  $\phi_X : \mathbb{R}^n \rightarrow \mathbb{C}$  définie par,  $\forall t \in \mathbb{R}^n$ ,  $\phi_X(t) = \mathbb{E}[e^{i\langle t, X \rangle}]$ . En particulier, si  $n = 1$ , on a  $\phi_X : t \mapsto \mathbb{E}[e^{itX}]$ .

*Exemple 2.3.9.* Pour tout  $t \in \mathbb{R}^n$ ,  $\phi_X(t) = \int_{\mathbb{R}^n} e^{i\langle t, x \rangle} d\mathbb{P}_X(x)$ . En particulier,  $\phi_X$  ne dépend que de  $\mathbb{P}_X$  et on parle aussi de la fonction caractéristique de la loi  $\mathbb{P}_X$ .

- Si  $X$  est discrète à valeur dans  $D$  dénombrable alors,  $\forall t \in \mathbb{R}^n$ ,  $\phi_X(t) = \sum_{x \in D} e^{i\langle t, x \rangle} \mathbb{P}(X = x)$ .
- Si  $X \sim \mathcal{X}(\lambda)$ , pour tout  $t \in \mathbb{R}$ ,  $\phi_X(t) = \sum_{k \in \mathbb{N}} e^{itk} e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k \in \mathbb{N}} \frac{(e^{it}\lambda)^k}{k!} = \exp(\lambda(e^{it} - 1))$ .
- Si  $X$  admet la densité  $f$  alors,  $\forall t \in \mathbb{R}^n$ ,  $\phi_X(t) = \int_{\mathbb{R}^n} e^{i\langle x, t \rangle} f(x) dx$ .
- Si  $X \sim \mathcal{N}(\mu, \sigma^2)$  alors  $\phi_X : t \mapsto e^{it\mu - \frac{t^2}{2}\sigma^2}$ . Exemple à connaître, preuve en TD.

**Lemme 2.3.10.** Soient  $a, b \in \mathbb{R}$  et  $X$  une v.a. réelle. Pour tout  $t \in \mathbb{R}$ , on a  $\phi_{aX+b}(t) = e^{itb} \phi_X(ta)$ .

*Démonstration.* Exercice. □

**Proposition 2.3.11.** La fonction  $\phi_X : \mathbb{R}^n \rightarrow \mathbb{C}$  est continue et,  $\forall t \in \mathbb{R}^n$ ,  $|\phi_X(t)| \leq \phi_X(0) = 1$ .

*Démonstration.* Pour tout  $t \in \mathbb{R}^n$ ,  $|\phi_X(t)| = |\mathbb{E}[e^{i\langle t, X \rangle}]| \leq \mathbb{E}[|e^{i\langle t, X \rangle}|] = \mathbb{E}[1] = 1 = \phi_X(0)$ . On admet la continuité dans le cas général et va la prouver dans un cas particulier.

Si  $X$  est discrète à valeurs dans  $\mathbb{N}$ , pour tout  $t \in \mathbb{R}$ , on a  $\phi_X(t) = \mathbb{E}[e^{itX}] = \sum_{k \in \mathbb{N}} p_k e^{itk}$  où,  $\forall k \in \mathbb{N}$ ,  $p_k = \mathbb{P}(X = k)$ . Soit  $k \in \mathbb{N}$ , on note  $f_k : t \mapsto p_k e^{itk}$  qui est continue de  $\mathbb{R}$  dans  $\mathbb{C}$  et telle que  $\|f_k\|_\infty = p_k$ . Comme  $\sum_{k \in \mathbb{N}} p_k = 1$ , on a  $\phi_X = \sum_{k \in \mathbb{N}} f_k$  avec convergence normale. Donc  $\phi_X$  est continue comme somme d'une série normalement convergente de fonctions continues. □

**Proposition 2.3.12.** Soient  $X$  une v.a.r. et  $p \in \mathbb{N}^*$ . Si  $X \in L^p(\Omega)$  alors  $\phi_X : \mathbb{R} \rightarrow \mathbb{C}$  est  $\mathcal{C}^p$  et, pour tout  $q \in \llbracket 0, p \rrbracket$ ,  $\phi_X^{(q)} : t \mapsto \mathbb{E}[(iX)^q e^{itX}]$ . En particulier,  $\phi_X^{(q)}(0) = i^q \mathbb{E}[X^q]$  pour tout  $q \in \llbracket 0, p \rrbracket$ .

*Démonstration dans un cas particulier.* On fait la preuve pour  $X$  est discrète à valeurs dans  $\mathbb{N}$ , avec les mêmes notations que précédemment. Pour tout  $k \in \mathbb{N}$ ,  $f_k : \mathbb{R} \rightarrow \mathbb{C}$  est  $\mathcal{C}^\infty$  et, pour tout  $q \in \mathbb{N}$ ,  $f_k^{(q)} : t \mapsto (ik)^q p_k e^{itk}$ . Ainsi  $\|f_k^{(q)}\|_\infty = k^q p_k$ , et  $\sum_{k \in \mathbb{N}} \|f_k^{(q)}\|_\infty = \sum_{k \in \mathbb{N}} k^q p_k = \mathbb{E}[|X|^q]$ . Si  $q \leq p$  alors  $X \in L^p(\Omega) \subset L^q(\Omega)$  et la série de fonction  $\sum_{k \in \mathbb{N}} f_k^{(q)}$  est normalement convergente sur  $\mathbb{R}$ .

On sait que  $\phi_X = \sum_{k \in \mathbb{N}} f_k$ . Soit  $q \in \llbracket 0, p-1 \rrbracket$ , si  $\phi_X$  est  $\mathcal{C}^q$  et  $\phi_X^{(q)} = \sum_{k \in \mathbb{N}} f_k^{(q)}$  alors, comme  $\sum_{k \in \mathbb{N}} f_k^{(q+1)}$  converge normalement,  $\phi_X$  est  $\mathcal{C}^{(q+1)}$  et  $\phi_X^{(q+1)} = \sum_{k \in \mathbb{N}} f_k^{(q+1)}$ . Par récurrence,  $\phi_X$  est de classe  $\mathcal{C}^p$  et, pour tout  $q \in \llbracket 0, p \rrbracket$ ,  $\phi_X^{(q)} : t \mapsto \sum_{k \in \mathbb{N}} f_k^{(q)}(t) = \sum_{k \in \mathbb{N}} (ik)^q e^{itk} p_k = \mathbb{E}[(iX)^q e^{itX}]$ . □

**Théorème 2.3.13.** Soient  $X$  et  $Y$  deux v.a. à valeurs dans  $\mathbb{R}^n$ . Si  $\phi_X = \phi_Y$  alors  $X$  et  $Y$  ont même loi. C'est-à-dire, la fonction caractéristique caractérise totalement la loi.

*Démonstration.* Admis. Voir [3, chap. 12.2]. □

**Proposition 2.3.14.** Soit  $X = (X_1, \dots, X_n)$  un vecteur aléatoire à valeurs dans  $\mathbb{R}^n$ , alors les  $(X_j)_{1 \leq j \leq n}$  sont indépendants si et seulement si  $\phi_X = \bigotimes_{j=1}^n \phi_{X_j}$ .

*Démonstration.* Si les  $(X_j)_{1 \leq j \leq n}$  sont indépendants, pour tout  $(t_1, \dots, t_n) \in \mathbb{R}^n$ , on a

$$\phi_X(t_1, \dots, t_n) = \mathbb{E} \left[ e^{i \sum_{j=1}^n t_j X_j} \right] = \mathbb{E} \left[ \prod_{j=1}^n e^{it_j X_j} \right] = \prod_{j=1}^n \mathbb{E} [e^{it_j X_j}] = \prod_{j=1}^n \phi_{X_j}(t_j).$$

Inversement, supposons  $\phi_X = \bigotimes_{j=1}^n \phi_{X_j}$ . Il existe  $Y_1, \dots, Y_n$  indépendants tels que,  $\forall j \in \llbracket 1, n \rrbracket$ ,  $d\mathbb{P}_{Y_j} = d\mathbb{P}_{X_j}$ . Soit  $Y = (Y_1, \dots, Y_n)$ , le sens direct montre que  $\phi_X = \bigotimes_{j=1}^n \phi_{X_j} = \bigotimes_{j=1}^n \phi_{Y_j} = \phi_Y$ . Par le théorème 2.3.13,  $\mathbb{P}_X = \mathbb{P}_Y = \bigotimes_{j=1}^n \mathbb{P}_{Y_j} = \bigotimes_{j=1}^n \mathbb{P}_{X_j}$ , donc  $X_1, \dots, X_n$  sont indépendants.  $\square$

**Lemme 2.3.15.** Soient  $X_1, \dots, X_n$  des v.a. indépendantes et  $S = \sum_{j=1}^n X_j$ , on a  $\phi_S = \prod_{j=1}^n \phi_{X_j}$ .

*Démonstration.* Exercice.  $\square$

### 2.3.3 Théorème de Lévy et théorème central limite

**Proposition 2.3.16.** Soient  $X$  et  $(X_k)_{k \in \mathbb{N}}$  des v.a. à valeurs dans  $\mathbb{R}^n$ , soient  $\phi$  et  $(\phi_k)_{k \in \mathbb{N}}$  leurs fonctions caractéristiques. On a  $X_k \xrightarrow[k \rightarrow +\infty]{\text{loi}} X$  si et seulement si,  $\forall t \in \mathbb{R}^n$ ,  $\phi_k(t) \xrightarrow[k \rightarrow +\infty]{} \phi(t)$ .

*Démonstration.* Supposons que  $X_k \xrightarrow[k \rightarrow +\infty]{\text{loi}} X$ . Soient  $t \in \mathbb{R}^n$ , on pose  $f_t : x \mapsto \cos(\langle t, x \rangle)$  et  $g_t : x \mapsto \sin(\langle t, x \rangle)$ . Comme  $f_t$  et  $g_t \in \mathcal{C}_b^0(\mathbb{R}^n)$  on a

$$\phi_k(t) = \mathbb{E} \left[ e^{i \langle t, X_k \rangle} \right] = \mathbb{E} [f_t(X_k)] + i \mathbb{E} [g_t(X_k)] \xrightarrow[k \rightarrow +\infty]{} \mathbb{E} [f_t(X)] + i \mathbb{E} [g_t(X)] = \mathbb{E} \left[ e^{i \langle t, X \rangle} \right] = \phi(t).$$

La réciproque est une conséquence du théorème de Lévy 2.3.17 ci-dessous.  $\square$

**Théorème 2.3.17** (Lévy). Soient  $(X_k)_{k \in \mathbb{N}}$  des v.a. à valeurs dans  $\mathbb{R}^n$ . Si leurs fonctions caractéristiques  $(\phi_k)_{k \in \mathbb{N}}$  convergent simplement vers une fonction  $\phi : \mathbb{R}^n \rightarrow \mathbb{C}$  continue en 0 alors  $\phi$  est la fonction caractéristique d'une v.a.  $X$  et  $X_k \xrightarrow[k \rightarrow +\infty]{\text{loi}} X$ .

*Démonstration.* Admis. Voir [3, chap. 14.2].  $\square$

**Théorème 2.3.18** (Théorème Central Limite). Soit  $(X_k)_{k \in \mathbb{N}^*}$  une suite de v.a.i.i.d. réelles  $L^2$  d'espérance commune  $\mu \in \mathbb{R}$  et de variance commune  $\sigma^2$ , où  $\sigma > 0$ . Pour tout  $n \in \mathbb{N}^*$ , on définit  $Y_n = \frac{1}{\sqrt{n}} \sum_{k=1}^n \frac{X_k - \mu}{\sigma} = \frac{\sqrt{n}}{\sigma} \left( \frac{1}{n} \sum_{k=1}^n X_k - \mu \right)$ . Alors  $Y_n \xrightarrow[n \rightarrow +\infty]{\text{loi}} \mathcal{N}(0, 1)$ .

*Démonstration.* Pour tout  $k \in \mathbb{N}^*$ , notons  $Z_k = \frac{X_k - \mu}{\sigma}$ . Sous les hypothèses du théorème, les  $(Z_k)_{k \in \mathbb{N}}$  sont des v.a.i.i.d. réelles  $L^2$  centrées et réduites. On note  $\phi : \mathbb{R} \rightarrow \mathbb{C}$  leur fonction caractéristique. Soient  $n \in \mathbb{N}^*$  et  $\phi_n$  la fonction caractéristique de  $Y_n = \frac{1}{\sqrt{n}} \sum_{k=1}^n Z_k$ . On va montrer que  $\phi_n$  converge simplement sur  $\mathbb{R}$  vers la fonction caractéristique  $t \mapsto e^{-\frac{t^2}{2}}$  de la loi  $\mathcal{N}(0, 1)$ . La proposition 2.3.16 donnera alors le résultat.

Soit  $t \in \mathbb{R}$ . Les  $(Z_k)_{k \geq 1}$  étant indépendantes, pour tout  $n \in \mathbb{N}^*$ , on a

$$\phi_n(t) = \phi_{\sum_{k=1}^n Z_k} \left( \frac{t}{\sqrt{n}} \right) = \prod_{k=1}^n \phi_{Z_k} \left( \frac{t}{\sqrt{n}} \right) = \left( \phi \left( \frac{t}{\sqrt{n}} \right) \right)^n.$$

D'après la proposition 2.3.12, la fonction  $\phi$  est  $\mathcal{C}^2$  et, par Taylor-Young,

$$\phi \left( \frac{t}{\sqrt{n}} \right) = \phi(0) + \frac{t}{\sqrt{n}} \phi'(0) + \frac{t^2}{2n} \phi''(0) + o \left( \frac{t^2}{n} \right) = 1 + \frac{it}{\sqrt{n}} \mathbb{E}[Z_1] - \frac{t^2}{2n} \mathbb{E}[Z_1^2] + o \left( \frac{1}{n} \right) = 1 - \frac{t^2}{2n} + o \left( \frac{1}{n} \right).$$

Donc  $\left( \phi \left( \frac{t}{\sqrt{n}} \right) \right)^n = \left( 1 - \frac{t^2}{2n} + o \left( \frac{1}{n} \right) \right)^n = \exp \left( n \log \left( 1 - \frac{t^2}{2n} + o \left( \frac{1}{n} \right) \right) \right) = e^{-\frac{t^2}{2} + o(1)} \xrightarrow[n \rightarrow +\infty]{} e^{-\frac{t^2}{2}}$ .  $\square$

# Chapitre 3

## Statistiques

### 3.1 Le point de vue des statistiques

Les probabilités et les statistiques manipulent les mêmes objets : espaces probabilisés, variables aléatoires, lois, ... mais avec des points de vue différents. Schématiquement, en probas, on travaille à partir d'une loi de probabilité connue, et on cherche à prédire le comportement des réalisations d'une variable aléatoire distribuée selon cette loi (en moyenne, avec grande probabilité, ...). En stats, le point de vue est inversé. On travaille à partir d'un jeu de données  $(x_1, \dots, x_n)$  que l'on suppose être des réalisations de variables aléatoires  $(X_1, \dots, X_n)$ , disons indépendantes et de même loi  $\mathbb{P}_X$  inconnue. On cherche alors à retrouver  $\mathbb{P}_X$  à partir des observations  $(x_1, \dots, x_n)$ . Plus réalistement, on cherche à estimer des grandeurs numériques associées à  $\mathbb{P}_X$  (par exemple son espérance) et à quantifier l'erreur commise par notre estimation.

#### 3.1.1 Exemple introductif

Un restaurateur rennais veut savoir s'il serait rentable d'investir dans une terrasse. Pour cela, il veut connaître le nombre  $X$  de jours de beau temps qu'il peut espérer sur une année.

Pour répondre à cette question, on peut poser un modèle probabiliste simpliste. On suppose que chaque jour il fait beau avec probabilité  $\theta \in [0; 1]$  et que la météo est indépendante d'un jour à l'autre. On modélise alors  $X$  comme la somme de  $N = 365$  variables de Bernoulli  $\mathcal{B}(\theta)$  indépendantes, c'est-à-dire  $X \sim \mathcal{B}(N, \theta)$ . En faisant des probas, on obtient  $\mathbb{E}[X] = N\theta$  et  $\text{Var}(X) = N\theta(1 - \theta)$  puis, par l'inégalité de Bienaymé–Chebishev, pour tout  $\varepsilon > 0$ ,

$$\mathbb{P}(|X - N\theta| \geq \varepsilon) \leq \frac{N\theta(1 - \theta)}{\varepsilon^2}.$$

Par exemple, pour  $\varepsilon = \sqrt{10N\theta(1 - \theta)}$  on obtient  $X \in [N\theta - \sqrt{10N\theta(1 - \theta)}; N\theta + \sqrt{10N\theta(1 - \theta)}]$  avec probabilité supérieure à 90%. Jusqu'ici on a supposé  $N$  et  $\theta$  connus et on en a déduit une information sur  $X$ . Détail gênant, on connaît  $N$  mais on n'a aucune idée de la valeur de  $\theta$ ...

C'est là que les stats entre en scène. Pour tout  $j \in \llbracket 1, 25 \rrbracket$ , on connaît le nombre  $x_j \in \llbracket 0, N \rrbracket$  de jours de beau temps à Rennes pour l'année 2000 +  $j$ . On suppose que le nombre de jours de beau temps est de même loi chaque année, et indépendant d'une année sur l'autre. Dans notre modèle, les  $(x_j)_{1 \leq j \leq 25}$  sont donc des réalisations de variables  $\mathcal{B}(N, \theta)$  indépendantes. Plus précisément, il existe un espace de probabilité  $(\Omega, \mathcal{A}, \mathbb{P})$ , des v.a.i.i.d.  $X_1, \dots, X_{25}$  de loi  $\mathcal{B}(N, \theta)$  définies sur cet espace, et  $\omega_0 \in \Omega$  tel que  $x_j = X_j(\omega_0)$  pour tout  $j \in \llbracket 1, 25 \rrbracket$ .

Considérons la variable aléatoire réelle  $\bar{X} = \frac{1}{25} \sum_{j=1}^{25} X_j$ . On a

$$\mathbb{E}[\bar{X}] = \frac{1}{25} \sum_{j=1}^{25} \mathbb{E}[X_j] = \mathbb{E}[X] = N\theta,$$

et, par indépendance,

$$\text{Var}(\bar{X}) = \frac{1}{(25)^2} \sum_{j=1}^{25} \text{Var}(X_j) = \frac{1}{25} \text{Var}(X) = \frac{N\theta(1-\theta)}{25}.$$

Par Bienaymé–Chebishev de nouveau, pour tout  $\varepsilon > 0$ ,

$$\mathbb{P}\left(\left|\frac{\bar{X}}{N} - \theta\right| \geq \varepsilon\right) = \mathbb{P}(|\bar{X} - N\theta| \geq N\varepsilon) \leq \frac{\text{Var}(\bar{X})}{(N\varepsilon)^2} = \frac{\theta(1-\theta)}{25N\varepsilon^2} \leq \frac{1}{100N\varepsilon^2},$$

où on a utilisé que le maximum de  $\theta \mapsto \theta(1-\theta)$  sur  $[0; 1]$  est atteint en  $\frac{1}{2}$  et vaut  $\frac{1}{4}$ . Par exemple, pour  $\varepsilon = \frac{1}{\sqrt{10N}}$  on obtient que  $\theta \in \left[\frac{\bar{X}}{N} - \frac{1}{\sqrt{10N}}; \frac{\bar{X}}{N} + \frac{1}{\sqrt{10N}}\right]$  avec probabilité supérieure à 90%. En notant  $\bar{x} = \frac{1}{25} \sum_{j=1}^{25} x_j = \bar{X}(\omega_0)$ , on peut estimer que  $\theta \in \left[\frac{\bar{x}}{N} - \frac{1}{\sqrt{10N}}; \frac{\bar{x}}{N} + \frac{1}{\sqrt{10N}}\right]$  et ré-injecter cette information dans la partie proba. On parle alors d'estimation *par intervalle de confiance*. Alternativement, on peut utiliser l'estimation *ponctuelle*  $\theta \simeq \frac{\bar{x}}{N}$ .

*Remarque 3.1.1.* • Dans la partie proba, on estime la probabilité que la grandeur aléatoire  $X$  appartienne à un intervalle déterministe. Dans la partie stats, on estime la probabilité que la grandeur déterministe  $\theta$  appartienne à un intervalle aléatoire.

- On serait tenté de dire que  $\theta \in \left[\frac{\bar{x}}{N} - \frac{1}{\sqrt{10N}}; \frac{\bar{x}}{N} + \frac{1}{\sqrt{10N}}\right]$  avec probabilité supérieure à 90%, mais ça n'a pas de sens. À ce stade tout est déterministe, ou bien  $\left[\frac{\bar{x}}{N} - \frac{1}{\sqrt{10N}}; \frac{\bar{x}}{N} + \frac{1}{\sqrt{10N}}\right]$  contient  $\theta$  ou bien il ne le contient pas.

Pour dire quelque chose du monde réel, on doit passer par une phase de modélisation lors de laquelle on fait de nombreuses hypothèses : existence d'une loi cachée ( $\mathbb{P}_X$ ), forme de cette loi ( $\mathcal{B}(N, \theta)$ , avec  $\theta \in [0; 1]$ ), indépendance des réalisations (plus exactement des v.a.  $X_1, \dots, X_{25}$ ), ... Ces hypothèses de modélisation sont toujours discutables et souvent invérifiables. Lorsqu'on pose un modèle statistique, il faut trouver un compromis entre :

- un modèle suffisamment complexe pour être fidèle au phénomène observé ;
- un modèle assez simple pour être étudié mathématiquement et fournir un résultat.

### 3.1.2 Vocabulaire des statistiques

**Définition 3.1.2** (Modèle statistique). Définir un *modèle statistique (paramétrique)* c'est se donner :

- un espace mesuré  $(\mathcal{X}, \mathcal{B})$ , dans lequel vivent nos observations,
- un *ensemble de paramètres*  $\Theta$ ,
- une famille  $(\mathbb{P}_\theta)_{\theta \in \Theta}$  de loi de probabilité sur  $(\mathcal{X}, \mathcal{B})$ .

Lorsque  $\Theta \subset \mathbb{R}^d$  pour un certain  $d \in \mathbb{N}^*$ , on dit que *le paramètre  $\theta$  est de dimension  $d$* .

Dans la plupart des cas concrets, on sera dans l'une des deux situations suivantes :  $\mathcal{X}$  (fini ou dénombrable) et  $\mathcal{B} = \mathcal{P}(\mathcal{X})$  ou  $\mathcal{X} = \mathbb{R}$  et  $\mathcal{B} = \mathcal{B}(\mathbb{R})$ .

*Exemple 3.1.3.* • Dans l'exemple introductif,  $\mathcal{X} = \llbracket 0, 365 \rrbracket$ ,  $\mathcal{B} = \mathcal{P}(\mathcal{X})$ ,  $\Theta = [0; 1]$  et, pour tout  $\theta \in [0; 1]$ ,  $\mathbb{P}_\theta = \mathcal{B}(N, \theta)$ . Notamment, le paramètre est de dimension 1.

- $(\mathcal{X}, \mathcal{B}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$  avec  $\Theta = \mathbb{R} \times \mathbb{R}_+^*$  et, pour tout  $\theta = (\mu, \sigma) \in \Theta$ ,  $\mathbb{P}_\theta = \mathcal{N}(\mu, \sigma^2)$ . Dans ce cas, le paramètre est de dimension 2.

**Définition 3.1.4** (Échantillon). Étant donné un modèle statistique  $(\mathcal{X}, \mathcal{B}, (\mathbb{P}_\theta)_{\theta \in \Theta})$  et  $n \in \mathbb{N}^*$ , un *échantillon (aléatoire de taille  $n$ )* ou  *$n$ -échantillon* est un  $n$ -uplet  $(X_1, \dots, X_n)$  de v.a.i.i.d. de loi  $\mathbb{P}_\theta$ , où  $\theta \in \Theta$  est fixé mais inconnu. La loi  $\mathbb{P}_\theta$  est appelée *loi-mère* de l'échantillon.

Un échantillon  $(X_1, \dots, X_n)$  de loi-mère  $\mathbb{P}_\theta$  est une variable aléatoire, à valeurs dans  $(\mathcal{X}^n, \mathcal{P}(\mathcal{X}^n))$  si  $\mathcal{X}$  est dénombrable ou dans  $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$  si  $(\mathcal{X}, \mathcal{B}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ , de loi  $\mathbb{P}_\theta^{\otimes n} := \mathbb{P}_\theta \otimes \dots \otimes \mathbb{P}_\theta$ . Un jeu d'observations est un élément  $(x_1, \dots, x_n) \in \mathcal{X}^n$ , pensé comme une réalisation de  $(X_1, \dots, X_n)$ , c'est-à-dire  $(x_1, \dots, x_n) = (X_1(\omega_0), \dots, X_n(\omega_0))$  pour un certain  $\omega_0 \in \Omega$ . Grossièrement, le but des statistiques est de deviner qui est  $\theta$  à partir des observations  $(x_1, \dots, x_n)$ .

**Définition 3.1.5** (Statistique). Soit  $(X_1, \dots, X_n)$  un échantillon issu d'un modèle  $(\mathcal{X}, \mathcal{B}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ , on appelle *statistique* toute variable aléatoire de la forme  $T = h(X_1, \dots, X_n)$ , où  $h$  est une fonction sur  $\mathcal{X}^n$  suffisamment gentille (par exemple continue) pour que  $T$  soit mesurable.

*Exemple 3.1.6.* Soient  $n \in \mathbb{N}^*$  et  $(X_1, \dots, X_n)$  un  $n$ -échantillon de v.a. réelles, issu d'un certain modèle statistique. On peut définir les statistiques classiques suivantes :

- la *moyenne empirique*  $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$  ;
- la *variance empirique*  $\tilde{S}_n = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2$  ;
- la *variance de l'échantillon*  $S_n = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X}_n)^2$  ;
- le maximum  $M_n = \max_{1 \leq k \leq n} X_k$ .

Connaissant la loi-mère de l'échantillon  $(X_1, \dots, X_n)$ , on ne sait en général pas déterminer la loi des statistiques associées. Dans les cas simples, on sait quand même dire des choses.

**Lemme 3.1.7.** Soient  $X$  une v.a.r. et  $(X_1, \dots, X_n)$  un  $n$ -échantillon dont la loi-mère est  $\mathbb{P}_X$ .

- Si  $X$  est  $L^1$  et  $\mathbb{E}[X] = \mu$  alors  $\bar{X}_n$  est  $L^1$  et  $\mathbb{E}[\bar{X}_n] = \mu$ .
- Si  $X$  est  $L^2$  et  $\text{Var}(X) = \sigma^2$  alors  $\bar{X}_n$  est  $L^2$  et  $\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$ .
- Si  $\phi_X$  est la fonction caractéristique de  $X$ , alors celle de  $\bar{X}_n$  est  $\phi_{\bar{X}_n} : t \mapsto \phi_X\left(\frac{t}{n}\right)^n$ .

*Démonstration.* Exercice. □

**Corollaire 3.1.8.** Si l'échantillon  $(X_1, \dots, X_n)$  est de loi-mère  $\mathcal{N}(\mu, \sigma^2)$  alors  $\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ .

*Démonstration.* Soit  $X \sim \mathcal{N}(\mu, \sigma^2)$ . D'après le lemme 3.1.7, pour tout  $t \in \mathbb{R}$  on a

$$\phi_{\bar{X}_n}(t) = \phi_X\left(\frac{t}{n}\right)^n = \left(\exp\left(i\frac{t}{n}\mu - \frac{t^2}{2n^2}\sigma^2\right)\right)^n = \exp\left(it\mu - \frac{t^2}{2} \frac{\sigma^2}{n}\right).$$

Comme la fonction caractéristique détermine totalement la loi, on a bien  $\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ . □

**Lemme 3.1.9.** Soient  $X$  une v.a.r.  $L^2$  de variance  $\sigma^2$  et  $(X_1, \dots, X_n)$  un  $n$ -échantillon dont la loi-mère est  $\mathbb{P}_X$ . Alors  $\tilde{S}_n$  est  $L^1$  avec  $\mathbb{E}[\tilde{S}_n] = \frac{n-1}{n}\sigma^2$  et  $S_n$  est  $L^1$  avec  $\mathbb{E}[S_n] = \sigma^2$ .

*Démonstration.* Soit  $k \in \llbracket 1, n \rrbracket$ , on a  $X_k - \bar{X}_n = X_k - \frac{1}{n} \sum_{i=1}^n X_i = \frac{n-1}{n} X_k + \sum_{i \neq k} (-\frac{1}{n}) X_i$ . Par combinaison linéaire  $X_k - \bar{X}_n$  est  $L^2$ , donc  $(X_k - \bar{X}_n)^2$  est  $L^1$ , donc  $\tilde{S}_n$  et  $S_n$  sont  $L^1$ . Ensuite,

$$\mathbb{E}[X_k - \bar{X}_n] = \frac{n-1}{n} \mathbb{E}[X_k] - \sum_{i \neq k} \frac{1}{n} \mathbb{E}[X_i] = \left( \frac{n-1}{n} - (n-1) \frac{1}{n} \right) \mathbb{E}[X] = 0.$$

Donc  $\text{Var}(X_k - \bar{X}_n) = \mathbb{E}[(X_k - \bar{X}_n)^2]$  et, par indépendance, on a

$$\text{Var}(X_k - \bar{X}_n) = \text{Var}\left(\frac{n-1}{n} X_k\right) + \sum_{i \neq k} \text{Var}\left(-\frac{1}{n} X_i\right) = \left(\frac{(n-1)^2}{n^2} + \frac{n-1}{n^2}\right) \sigma^2 = \frac{n-1}{n} \sigma^2.$$

Donc  $\mathbb{E}[\tilde{S}_n] = \frac{1}{n} \sum_{k=1}^n \mathbb{E}[(X_k - \bar{X}_n)^2] = \frac{n-1}{n} \sigma^2$  et  $\mathbb{E}[S_n] = \mathbb{E}\left[\frac{n}{n-1} \tilde{S}_n\right] = \frac{n}{n-1} \mathbb{E}[\tilde{S}_n] = \sigma^2$ .  $\square$

**Lemme 3.1.10.** Soient  $X$  une v.a.r. de fonction de répartition  $F_X$  et  $(X_1, \dots, X_n)$  un  $n$ -échantillon dont la loi-mère est  $\mathbb{P}_X$ . Alors la fonction de répartition de  $M_n$  est  $F_{M_n} = (F_X)^n$ .

*Démonstration.* Exercice.  $\square$

## 3.2 Estimateurs ponctuels

On considère un modèle statistique  $(\mathcal{X}, \mathcal{B}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ . Lorsque  $(\mathbb{P}_\theta)_{\theta \in \Theta}$  est une famille de lois classiques, on notera parfois simplement  $(\mathbb{P}_\theta)_{\theta \in \Theta}$  pour le modèle, l'ensemble mesuré  $(\mathcal{X}, \mathcal{B})$  étant implicite. Dans l'exemple introductif, on peut prendre comme modèle  $(\mathbb{N}, \mathcal{P}(\mathbb{N}), (\mathcal{B}(N, \theta))_{\theta \in [0;1]})$  avec  $N = 365$ , ce qu'on abrégera en  $(\mathcal{B}(N, \theta))_{\theta \in [0;1]}$ .

### 3.2.1 Notion d'estimateur

**Définition 3.2.1** (Estimateur et estimation). Soient  $\psi$  une fonction sur  $\Theta$  et  $T = h(X_1, \dots, X_n)$  une statistique basée sur un  $n$ -échantillon dans notre modèle. Dire que  $T$  est un *estimateur* de  $\psi(\theta)$ , c'est affirmer qu'on pense que les réalisations de  $T$  sont proches de  $\psi(\theta)$ .

Si  $T$  est un estimateur de  $\psi(\theta)$ , soient  $\omega_0 \in \Omega$  et  $(x_1, \dots, x_n) = (X_1(\omega_0), \dots, X_n(\omega_0))$  une réalisation de notre  $n$ -échantillon, on dit que  $h(x_1, \dots, x_n) = T(\omega_0)$  est une *estimation* de  $\psi(\theta)$ .

*Remarque 3.2.2.* • Le terme estimateur est flou. Il n'y a pas de différence formelle entre estimateurs et statistiques. On ajoute juste une nuance disant que cette statistique est intéressante.

- "Les réalisations de  $T$  sont proches de  $\psi(\theta)$ " signifie que : pour tout  $\theta \in \Theta$ , si la loi-mère de l'échantillon est  $\mathbb{P}_\theta$ , alors la distribution de  $T$  (qui dépend de  $\theta$  via  $(X_1, \dots, X_n)$ ) "est concentrée" autour de  $\psi(\theta)$ . C'est toujours flou...
- Souvent  $\Theta \subset \mathbb{R}$  avec  $\psi : \theta \mapsto \theta$  l'inclusion, et on s'intéresse à des estimateurs de  $\theta$ .

*Exemple 3.2.3.* • Dans l'exemple introductif, le modèle est  $(\mathcal{B}(N, \theta))_{\theta \in \Theta}$  avec  $N = 365$ ,  $n = 25$ , et on a considéré la statistique  $T = \frac{1}{N} \bar{X} = \frac{1}{25N} \sum_{k=1}^{25} X_k$  comme un estimateur de  $\theta$ .

- Dans le modèle  $(\mathcal{N}(\mu, \sigma^2))_{(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+^*}$ , pour  $\psi : (\mu, \sigma) = \theta \mapsto \mu$  de  $\Theta = \mathbb{R} \times \mathbb{R}_+^*$  dans  $\mathbb{R}$ , la statistique  $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$  est un estimateur de  $\mu = \psi(\theta)$ .

### 3.2.2 Qualité d'un estimateur

N'importe quelle fonction mesurable  $h : \mathcal{X}^n \rightarrow \mathbb{R}$  appliquée à un  $n$ -échantillon  $(X_1, \dots, X_n)$  définit une statistique  $T = h(X_1, \dots, X_n)$ . On va introduire des notions permettant de quantifier la pertinence de  $T$  en tant qu'estimateur d'une quantité  $\psi(\theta)$ , où  $\psi : \Theta \rightarrow \mathbb{R}$ .

Dans la suite, pour tout  $\theta \in \Theta$ , on note  $\mathbb{E}_\theta[T]$  (resp.  $\text{Var}_\theta(T)$ ) l'espérance (resp. la variance) de  $T = h(X_1, \dots, X_n)$  lorsque la loi-mère de l'échantillon est  $\mathbb{P}_\theta$ , sous réserve qu'elle soit bien définie.

**Définition 3.2.4** (Biais). Soit  $\theta \in \Theta$  tel que  $\mathbb{E}_\theta[|T|] < +\infty$ . On définit le *biais*  $b_T(\theta) \in \mathbb{R}$  de  $T$  en  $\theta$  par  $b_T(\theta) = \mathbb{E}_\theta[T] - \psi(\theta)$ . On dit que  $T$  est un estimateur de  $\psi(\theta)$  *sans biais* si,  $\forall \theta \in \Theta$ ,  $b_T(\theta) = 0$ .

**Définition 3.2.5** (Risque quadratique). Soit  $\theta \in \Theta$  tel que  $\mathbb{E}_\theta[T^2] < +\infty$ . On définit le *risque* (quadratique)  $R_T(\theta) \geq 0$  de  $T$  en  $\theta$  par  $R_T(\theta) = \mathbb{E}_\theta[(T - \psi(\theta))^2]$ .

**Lemme 3.2.6.** Pour tout  $\theta \in \Theta$  tel que  $\mathbb{E}_\theta[T^2] < +\infty$ , on a  $R_T(\theta) = b_T(\theta)^2 + \text{Var}_\theta(T)$ .

*Démonstration.* On calcule

$$\begin{aligned} R_T(\theta) &= \mathbb{E}_\theta[(T - \psi(\theta))^2] = \mathbb{E}_\theta[(T - \mathbb{E}_\theta[T] + \mathbb{E}_\theta[T] - \psi(\theta))^2] = \mathbb{E}_\theta[(T - \mathbb{E}_\theta[T] + b_T(\theta))^2] \\ &= \mathbb{E}_\theta[(T - \mathbb{E}_\theta[T])^2] + \mathbb{E}_\theta[2(T - \mathbb{E}_\theta[T])b_T(\theta)] + \mathbb{E}_\theta[b_T(\theta)^2] \\ &= \text{Var}_\theta(T) + 2b_T(\theta)\mathbb{E}_\theta[(T - \mathbb{E}_\theta[T])] + b_T(\theta)^2 = \text{Var}_\theta(T) + b_T(\theta)^2. \quad \square \end{aligned}$$

*Exemple 3.2.7.* Dans le modèle  $(\mathcal{B}(N, \theta))_{\theta \in [0;1]}$  de l'exemple introductif, on a montré que, pour tout  $\theta \in [0;1]$ ,  $\mathbb{E}_\theta[\bar{X}] = N\theta$ , donc  $\mathbb{E}_\theta[\frac{1}{N}\bar{X}] = \theta$ . Donc  $T = \frac{1}{N}\bar{X}$  est un estimateur sans biais de  $\theta$ . Ensuite, pour tout  $\theta \in [0;1]$ ,

$$R_T(\theta) = \text{Var}_\theta(T) = \frac{1}{N^2} \text{Var}_\theta(\bar{X}) = \frac{\theta(1-\theta)}{25N} \leq \frac{1}{100N}.$$

**Lemme 3.2.8.** Soit  $\theta \in \Theta$  tel que  $\mathbb{E}_\theta[T^2] < +\infty$ . Pour tout  $\varepsilon > 0$  on a  $\mathbb{P}_\theta(|T - \psi(\theta)| > \varepsilon) \leq \frac{R_T(\theta)}{\varepsilon^2}$ , où l'indice  $\theta$  signifie qu'on calcule la probabilité sous l'hypothèse que la loi-mère est  $\mathbb{P}_\theta$ .

*Démonstration.* Soit  $\varepsilon > 0$ , par l'inégalité de Markov,

$$\mathbb{P}_\theta(|T - \psi(\theta)| > \varepsilon) = \mathbb{P}_\theta((T - \psi(\theta))^2 > \varepsilon^2) \leq \frac{1}{\varepsilon^2} \mathbb{E}_\theta[(T - \psi(\theta))^2] = \frac{1}{\varepsilon^2} R_T(\theta). \quad \square$$

Plus le risque  $R_T(\theta)$  est petit, plus la variable  $T$  est concentrée autour de  $\psi(\theta)$ . Notamment, si  $R_T(\theta) = 0$  alors  $T = \psi(\theta)$  presque sûrement. Dans ces estimations tout dépend de  $\theta$ , qui est inconnu. On cherche donc à avoir  $R_T(\theta)$  petit indépendamment de  $\theta$ , comme dans l'exemple introductif. Un bon estimateur est donc un estimateur de risque faible (uniformément en  $\theta$ ). On peut alors dire qu'un estimateur est meilleur qu'un autre si son risque est (uniformément) plus petit.

**Définition 3.2.9** (Domination). Soient  $T$  et  $\tilde{T}$  deux estimateurs d'une même quantité  $\psi(\theta)$ . On dit que  $T$  *domine*  $\tilde{T}$  si :

- pour tout  $\theta \in \Theta$ ,  $R_T(\theta) \leq R_{\tilde{T}}(\theta)$  ;
- il existe  $\theta_0 \in \Theta$  tel que  $R_T(\theta_0) < R_{\tilde{T}}(\theta_0)$ .

Jusqu'à présent, on a travaillé avec des échantillons de taille fixée. Pour les estimateurs usuels, tels la moyenne empirique  $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$ , on a en fait une suite d'estimateurs indexée par  $n \in \mathbb{N}^*$  et on peut en étudier le comportement asymptotique quand  $n \rightarrow +\infty$ .

Dans la suite, on considère une suite  $(X_k)_{k \in \mathbb{N}^*}$  de v.a.i.i.d. de loi-mère  $\mathbb{P}_\theta$  inconnue. On considère, pour tout  $n \in \mathbb{N}^*$ , un estimateur  $T_n = h_n(X_1, \dots, X_n)$  de  $\psi(\theta)$  basé sur le  $n$ -échantillon  $(X_1, \dots, X_n)$ .

Quand  $n$  augmente, on augmente la taille du jeu de données  $(x_1, \dots, x_n) = (X_1(\omega_0), \dots, X_n(\omega_0))$  sur lequel se base l'estimation  $\psi(\theta) \simeq h_n(x_1, \dots, x_n)$  associée à  $T_n$ . Comme on utilise plus d'information, il est logique d'espérer un meilleur résultat. À la limite  $n \rightarrow +\infty$ , on utilise moralement une infinité d'informations, et il est raisonnable d'espérer que ça suffise à reconstruire  $\psi(\theta)$  parfaitement. Un comportement souhaitable est donc que  $(T_n)_{n \in \mathbb{N}^*}$  converge vers  $\psi(\theta)$  en un certain sens.

**Définition 3.2.10** (Consistance). Soit  $(T_n)_{n \in \mathbb{N}^*}$  une suite d'estimateurs de  $\psi(\theta)$  comme ci-dessus.

- Si, pour tout  $\theta \in \Theta$ , sous l'hypothèse que la loi-mère est  $\mathbb{P}_\theta$  on a  $T_n \xrightarrow[n \rightarrow +\infty]{\mathbb{P}} \psi(\theta)$ , alors on dit que la suite  $(T_n)_{n \in \mathbb{N}^*}$  est *consistante*.
- Si, pour tout  $\theta \in \Theta$ , sous l'hypothèse que la loi-mère est  $\mathbb{P}_\theta$  on a  $T_n \xrightarrow[n \rightarrow +\infty]{\text{p.s.}} \psi(\theta)$ , alors on dit que la suite  $(T_n)_{n \in \mathbb{N}^*}$  est *fortement consistante*.

Par abus de langage, on dira que  $T_n$  est un *estimateur consistant* (resp. *fortement consistant*).

**Définition 3.2.11** (Normalité asymptotique). Soit  $(T_n)_{n \in \mathbb{N}^*}$  une suite d'estimateurs de  $\psi(\theta)$  comme ci-dessus. On dit que  $(T_n)_{n \in \mathbb{N}^*}$  est *asymptotiquement normale* s'il existe une suite  $(R_n)_{n \in \mathbb{N}^*}$  positive telle que : pour tout  $\theta \in \Theta$ , il existe  $\sigma_\theta > 0$  tel que, sous l'hypothèse que la loi-mère est  $\mathbb{P}_\theta$ , on a

$$R_n(T_n - \psi(\theta)) \xrightarrow[n \rightarrow +\infty]{\text{loi}} \mathcal{N}(0, \sigma_\theta^2).$$

*Exemple 3.2.12.* Revenons au modèle  $(\mathcal{B}(N, \theta))_{\theta \in [0;1]}$  de l'exemple introductif. Pour tout  $\theta \in [0;1]$ , les  $(X_k)_{k \in \mathbb{N}^*}$  sont des v.a.i.i.d.  $L^1$  d'espérance  $N\theta$ . Par la loi forte des grands nombres, on a donc  $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k \xrightarrow[n \rightarrow +\infty]{\text{p.s.}} N\theta$ . Donc, sous  $\mathbb{P}_\theta$ , on a  $T_n = \frac{1}{N} \bar{X}_n \xrightarrow[n \rightarrow +\infty]{\text{p.s.}} \theta$ . Donc  $(T_n)_{n \in \mathbb{N}^*}$  est une suite fortement consistante d'estimateurs de  $\theta$ .

Par ailleurs, si on suppose  $\theta \in ]0;1[$ , on a

$$(T_n - \theta) = \frac{1}{N} \left( \frac{1}{n} \sum_{k=1}^n X_k - N\theta \right) = \sqrt{\frac{\theta(1-\theta)}{Nn}} \left( \frac{1}{\sqrt{n}} \sum_{k=1}^n \frac{X_k - N\theta}{\sqrt{N\theta(1-\theta)}} \right).$$

Comme les  $(X_k)_{k \in \mathbb{N}^*}$  sont  $L^2$  d'espérance  $N\theta$  et de variance  $N\theta(1-\theta)$ , par le théorème central limite on a  $\frac{1}{\sqrt{n}} \sum_{k=1}^n \frac{X_k - N\theta}{\sqrt{N\theta(1-\theta)}} \xrightarrow[n \rightarrow +\infty]{\text{loi}} \mathcal{N}(0, 1)$ . On en déduit que

$$\sqrt{n}(T_n - \theta) = \sqrt{\frac{\theta(1-\theta)}{N}} \left( \frac{1}{\sqrt{n}} \sum_{k=1}^n \frac{X_k - N\theta}{\sqrt{N\theta(1-\theta)}} \right) \xrightarrow[n \rightarrow +\infty]{\text{loi}} \mathcal{N}\left(0, \frac{\theta(1-\theta)}{N}\right).$$

Si on restreint l'espace des paramètres à  $]0;1[$ , la suite  $(T_n)_{n \in \mathbb{N}^*}$  est donc asymptotiquement normale. Si  $\theta \in \{0;1\}$  alors, pour tout  $n \in \mathbb{N}^*$ ,  $T_n = \theta$  presque sûrement sous  $\mathbb{P}_\theta$ . Pour ces paramètres, on n'a aucune chance d'avoir une limite en loi non-nulle, même après renormalisation.

### 3.2.3 Méthode des moments

Supposons que notre modèle  $(\mathcal{X}, \mathcal{B}, (\mathbb{P}_\theta)_{\theta \in \Theta})$  est tel que  $\mathcal{X} \subset \mathbb{R}$  et  $\Theta \subset \mathbb{R}^d$ . On va présenter une méthode classique, appelée *méthode des moments*, pour construire des estimateurs de  $\theta$  ( $d = 1$ ) ou de ses composantes ( $d > 1$ ). Dans la suite on note  $X$  une v.a.r. de loi  $\mathbb{P}_\theta$ .

**Définition 3.2.13** (Fonctions moments). Soit  $p \in \mathbb{N}^*$ . Si pour tout  $\theta \in \Theta$  on a  $\mathbb{E}_\theta[|X|^p] < +\infty$ , on pose  $\mu_p : \theta \mapsto \mathbb{E}_\theta[X^p]$  de  $\Theta$  dans  $\mathbb{R}$ .

On présente la méthode dans le cas le plus simple, qui est aussi le plus utile en pratique. Supposons que, pour tout  $\theta \in \Theta$ , la v.a.  $X$  est  $L^1$ . Soit  $(X_k)_{k \in \mathbb{N}^*}$  une suite de v.a.i.i.d. de loi-mère  $\mathbb{P}_\theta$ . D'après la loi des grands nombres, pour tout  $\theta \in \Theta$ , on a

$$\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k \xrightarrow[n \rightarrow +\infty]{\text{p.s.}} \mathbb{E}_\theta[X] = \mu_1(\theta).$$

Il est donc raisonnable de considérer  $\bar{X}_n$  comme un estimateur de  $\mu_1(\theta)$ . Une réalisation  $\bar{x}_n$  de  $\bar{X}_n$  fournit alors une estimation  $\mu_1(\theta) \simeq \bar{x}_n$ . Si  $\mu_1$  est injective, elle réalise une bijection de  $\Theta$  vers son image, et un choix naturel est d'estimer  $\theta \simeq \mu_1^{-1}(\bar{x}_n)$ . L'estimateur de  $\theta$  correspondant est  $\mu_1^{-1}(\bar{X}_n)$ .

**Définition 3.2.14** (Estimateur par la méthode des moments, dimension 1). Supposons la fonction  $\mu_1 : \Theta \rightarrow \mathbb{R}$  bien définie et injective. Pour tout  $n \in \mathbb{N}^*$  tel que  $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$  est presque sûrement à valeurs dans  $\mu_1(\Theta)$ , on définit l'*estimateur de  $\theta$  par la méthode des moments* comme  $T_n = \mu_1^{-1}(\bar{X}_n)$ .

*Exemple 3.2.15.* Dans le modèle  $(\mathcal{E}(\theta))_{\theta > 0}$ , c'est-à-dire  $\Theta = ]0, +\infty[$  et  $\mathbb{P}_\theta$  est la loi exponentielle de paramètre  $\theta$  sur  $]0, +\infty[$ . Pour tout  $\theta > 0$ , on a  $\mu_1(\theta) = \mathbb{E}_\theta[X] = \frac{1}{\theta}$ . En particulier,  $\mu_1$  est injective d'image  $]0, +\infty[$ , et  $\mu_1^{-1} : x \mapsto \frac{1}{x}$  de  $]0, +\infty[$  dans lui-même. Sous la loi-mère  $\mathbb{P}_\theta$ , pour tout  $n \in \mathbb{N}^*$ ,  $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$  est à valeurs strictement positives, et  $T_n = \frac{1}{\bar{X}_n} = \frac{n}{\sum_{k=1}^n X_k}$  est l'estimateur de  $\theta$  par la méthode des moments.

*Remarque 3.2.16.* Dans les cas concrets,  $\mu_1$  est en général continue. Pour espérer que  $\mu_1(\theta) \simeq \bar{x}_n$  implique  $\theta \simeq \mu_1^{-1}(\bar{x}_n)$ , il faut que  $\mu_1^{-1}$  soit un minimum régulière. Sauf cas pathologiques, on a  $\mu_1^{-1}$  continue aussi, et donc  $\mu_1$  réalise un homéomorphisme de  $\Theta \subset \mathbb{R}^d$  vers  $\mu_1(\Theta) \subset \mathbb{R}$ . Si  $\Theta$  n'est pas d'intérieur vide, comme c'est en général le cas, cela impose  $d = 1$ . On ne peut donc espérer construire un estimateur raisonnable de  $\theta$  par cette méthode que si le paramètre est de dimension 1.

Il arrive que  $\mu_1$  ne soit pas injective, ou plus généralement pas sympathique. Si  $X$  est  $L^p$  sous  $\mathbb{P}_\theta$ , alors  $X^p$  est  $L^1$  et  $\frac{1}{n} \sum_{k=1}^n X_k^p \xrightarrow[n \rightarrow +\infty]{\text{p.s.}} \mathbb{E}_\theta[X^p] = \mu_p(\theta)$ . On peut alors raisonner de même avec  $\mu_p$ .

**Définition 3.2.17** (Estimateur par la méthode des moments, variante). Soit  $p \in \mathbb{N}^*$ , supposons la fonction  $\mu_p : \Theta \rightarrow \mathbb{R}$  bien définie et injective. Pour tout  $n \in \mathbb{N}^*$  tel que  $\frac{1}{n} \sum_{k=1}^n X_k^p$  est presque sûrement à valeurs dans  $\mu_p(\Theta)$ , on définit un estimateur de  $\theta$  par  $T_n = \mu_p^{-1}\left(\frac{1}{n} \sum_{k=1}^n X_k^p\right)$ .

De nouveau, si  $\mu_p$  et  $\Theta$  sont raisonnables, cette variante ne peut fonctionner que si le paramètre est de dimension 1. Pour un paramètre de dimension  $d \geq 2$ , il faut utiliser plusieurs fonctions moments simultanément. Si  $X$  est  $L^d$  sous  $\mathbb{P}_\theta$ , alors on a

$$\forall p \in \llbracket 1, d \rrbracket, \quad \frac{1}{n} \sum_{k=1}^n X_k^p \xrightarrow[n \rightarrow +\infty]{\text{p.s.}} \mathbb{E}_\theta[X^p] = \mu_p(\theta).$$

Si la fonction  $\mu_{\llbracket 1, d \rrbracket} : \theta \mapsto (\mu_1(\theta), \dots, \mu_d(\theta))$  est injective de  $\Theta$  vers  $\mathbb{R}^d$ , on joue au même jeu qu'avant.

**Définition 3.2.18** (Estimateur par la méthode des moments, dimension  $d$ ). Soit  $d \in \mathbb{N}^*$ , supposons  $\Theta \subset \mathbb{R}^d$  et la fonction  $\mu_{\llbracket 1, d \rrbracket} : \Theta \rightarrow \mathbb{R}^d$  bien définie et injective. Pour tout  $n \in \mathbb{N}^*$  tel que  $(\frac{1}{n} \sum_{k=1}^n X_k^p)_{1 \leq p \leq d}$  est presque sûrement à valeurs dans  $\mu_{\llbracket 1, d \rrbracket}(\Theta)$ , on définit un estimateur de  $\theta$  par

$$T_n = \mu_{\llbracket 1, d \rrbracket}^{-1} \left( \frac{1}{n} \sum_{k=1}^n X_k, \frac{1}{n} \sum_{k=1}^n X_k^2, \dots, \frac{1}{n} \sum_{k=1}^n X_k^d \right).$$

Par estimateur de  $\theta = (\theta_1, \dots, \theta_d)$ , on entend que  $T_n = (T_{n,1}, \dots, T_{n,d})$  est tel que, pour tout  $j \in \llbracket 1, d \rrbracket$ ,  $T_{n,j}$  est un estimateur de  $\theta_j$ .

**Théorème 3.2.19** (Consistance de la méthode des moments). *Supposons que :*

- $\Theta \subset \mathbb{R}$  est un intervalle ;
- $\mu_1 : \Theta \rightarrow \mathbb{R}$  est bien définie (i.e.  $\forall \theta \in \Theta, \mathbb{E}_\theta[|X|] < +\infty$ ), continue et injective ;
- pour tout  $n \in \mathbb{N}^*$ , l'estimateur  $T_n = \mu_1^{-1}(\bar{X}_n)$  de  $\theta$  est bien défini.

Alors  $(T_n)_{n \in \mathbb{N}^*}$  est une suite fortement consistante d'estimateurs de  $\theta$ .

*Démonstration.* Sous  $\mathbb{P}_\theta$ , la variable  $X$  est  $L^1$ . Par la loi forte des grands nombres, on a donc  $\bar{X}_n \xrightarrow[n \rightarrow +\infty]{\text{p.s.}} \mu_1(\theta)$ . Sous nos hypothèses,  $\mu_1$  est une bijection de  $\Theta$  vers  $\mu_1(\Theta)$ . Pour conclure, il suffit de vérifier que  $\mu_1^{-1}$  est continue. On aura alors  $T_n = \mu_1^{-1}(\bar{X}_n) \xrightarrow[n \rightarrow +\infty]{\text{p.s.}} \theta$ .

Comme  $\mu_1$  est continue, injective et définie sur un intervalle, elle est strictement monotone (faire un dessin). On conclut en utilisant le résultat classique qu'une fonction continue et strictement monotone réalise un homéomorphisme vers son image. En particulier,  $\mu_1^{-1}$  est continue.  $\square$

*Remarque 3.2.20.* Ce résultat s'étend à la variante, en remplaçant  $\mu_1$  par  $\mu_p$  et  $\bar{X}_n$  par  $\frac{1}{n} \sum_{k=1}^n X_k^p$ . Il s'étend aussi en dimension  $d$ , en remplaçant  $\mu_1$  par  $\mu_{\llbracket 1, d \rrbracket}$  et en supposant  $\Theta \subset \mathbb{R}^d$  ouvert. En pratique,  $\mu_1^{-1}$  est explicite, et il est souvent plus simple de refaire le raisonnement plutôt que d'appliquer le théorème 3.2.19.

**Théorème 3.2.21** (Normalité asymptotique de la méthode des moments). *On suppose que :*

- $\Theta \subset \mathbb{R}$  est un intervalle ouvert ;
- $\forall \theta \in \Theta, \mathbb{E}_\theta[X^2] < +\infty$  ;
- $\mu_1$  est de classe  $\mathcal{C}^1$  et  $\mu_1'$  ne s'annule pas ;
- pour tout  $n \in \mathbb{N}^*$ , l'estimateur  $T_n = \mu_1^{-1}(\bar{X}_n)$  de  $\theta$  est bien défini.

Alors, pour tout  $\theta \in \Theta$ , on a  $\sqrt{n}(T_n - \theta) \xrightarrow[n \rightarrow +\infty]{\text{loi}} \mathcal{N}\left(0, \frac{\text{Var}_\theta(X)}{\mu_1'(\theta)^2}\right)$ .

*Démonstration.* Admis.  $\square$

### 3.2.4 Méthode du maximum de vraisemblance

On va présenter une seconde méthode pour construire des estimateurs, pour une famille de lois discrètes ou à densité. On considère un modèle statistique  $(\mathcal{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ , où  $\mathcal{X} \subset \mathbb{R}$  et  $\Theta \subset \mathbb{R}^d$ .

Commençons par le cas à densité. On suppose que, pour tout  $\theta \in \Theta$ , la loi  $\mathbb{P}_\theta$  est une mesure de probabilité à densité sur  $\mathcal{X} \subset \mathbb{R}$ . On définit alors  $f : \Theta \times \mathcal{X} \rightarrow \mathbb{R}_+$  telle que, pour tout  $\theta \in \Theta$ , la fonction  $f(\theta, \cdot) : x \mapsto f(\theta, x)$  est la densité de  $\mathbb{P}_\theta$ . Soit  $(X_1, \dots, X_n)$  un  $n$ -échantillon de loi-mère  $\mathbb{P}_\theta$ . Ce vecteur aléatoire admet la densité  $(x_1, \dots, x_n) \mapsto \prod_{k=1}^n f(\theta, x_k)$ .

**Définition 3.2.22** (Fonction de vraisemblance, cas à densité). On définit  $f_n : \Theta \times \mathcal{X}^n \rightarrow \mathbb{R}_+$  par  $f_n : (\theta, x_1, \dots, x_n) \mapsto \prod_{k=1}^n f(\theta, x_k)$ . On appelle *fonction de vraisemblance* en  $(x_1, \dots, x_n) \in \mathcal{X}^n$  la fonction  $V_{(x_1, \dots, x_n)} : \theta \mapsto f_n(\theta, x_1, \dots, x_n)$  de  $\Theta$  dans  $\mathbb{R}_+$ .

Le principe de la méthode est que, si on connaît une réalisation  $(x_1, \dots, x_n) \in \mathcal{X}^n$  de l'échantillon, alors le paramètre le plus plausible est celui pour lequel la fonction de vraisemblance  $V_{(x_1, \dots, x_n)}$  est maximale, c'est-à-dire  $\theta_0 \in \Theta$  tel que  $V_{(x_1, \dots, x_n)}(\theta_0) = \max_{\theta \in \Theta} V_{(x_1, \dots, x_n)}(\theta)$ . Pour que cela ait du sens, il faut que  $V_{(x_1, \dots, x_n)}$  atteigne son maximum en un unique point. On définit  $D_n \subset \mathcal{X}^n \subset \mathbb{R}^n$  l'ensemble des points  $(x_1, \dots, x_n)$  pour lesquels c'est le cas.

**Définition 3.2.23** (Argmax). Soit  $V : \Theta \rightarrow \mathbb{R}$ . Si  $V$  atteint un maximum en un unique point  $\theta_0 \in \Theta$ , on dit que  $\theta_0$  est l'*argmax* de  $V$ , et on le note  $\theta_0 = \arg \max_{\theta \in \Theta} V(\theta)$ .

**Définition 3.2.24** (Estimateur du maximum de vraisemblance, cas à densité). Soit  $h_n : D_n \rightarrow \Theta$  la fonction  $h_n : (x_1, \dots, x_n) \mapsto \arg \max_{\theta \in \Theta} V_{(x_1, \dots, x_n)}(\theta)$ . Si l'échantillon  $(X_1, \dots, X_n)$  est presque sûrement à valeurs dans  $D_n$ , l'*estimateur du maximum de vraisemblance* de  $\theta$  est  $T_n = h_n(X_1, \dots, X_n)$ .

*Exemple 3.2.25.* Dans le modèle  $(\mathcal{E}(\theta))_{\theta > 0}$ , on a  $\Theta = ]0; +\infty[$ , et on peut prendre  $\mathcal{X} = ]0; +\infty[$ . Pour tout  $\theta > 0$ , la loi  $\mathbb{P}_\theta$  admet la densité  $x \mapsto \theta e^{-\theta x}$ . On a donc  $f : (\theta, x) \mapsto \theta e^{-\theta x}$  et

$$f_n : (\theta, x_1, \dots, x_n) \mapsto \prod_{k=1}^n \theta e^{-\theta x_k} = \theta^n e^{-\theta \sum_{k=1}^n x_k}.$$

Soit  $(x_1, \dots, x_n) \in \mathcal{X}^n$ . La fonction  $V_{(x_1, \dots, x_n)} : \theta \mapsto \theta^n \exp(-\theta \sum_{k=1}^n x_k)$  est continue, strictement positive et elle tend vers 0 lorsque  $\theta \rightarrow 0$  (car  $n \in \mathbb{N}^*$ ) et lorsque  $\theta \rightarrow +\infty$  (car  $\sum_{k=1}^n x_k > 0$ ). Elle atteint donc un maximum sur  $\Theta = ]0; +\infty[$ . Cette fonction étant en fait dérivable, ce maximum est atteint en un point d'annulation de sa dérivée  $V'_{(x_1, \dots, x_n)} : \theta \mapsto (n - \theta \sum_{k=1}^n x_k) \theta^{n-1} e^{-\theta \sum_{k=1}^n x_k}$ . Or,

$$V'_{(x_1, \dots, x_n)}(\theta) = 0 \iff n - \theta \sum_{k=1}^n x_k = 0 \iff \theta = \frac{n}{\sum_{k=1}^n x_k}.$$

Donc  $(x_1, \dots, x_n) \in D_n$  et  $\arg \max_{\theta > 0} V_{(x_1, \dots, x_n)}(\theta) = \frac{n}{\sum_{k=1}^n x_k}$ . Ainsi,  $h_n : (x_1, \dots, x_n) \mapsto \frac{n}{\sum_{k=1}^n x_k}$  de  $D_n = \mathcal{X}^n$  vers  $\Theta$ . Comme  $\mathbb{P}((X_1, \dots, X_n) \in D_n) = 1$ , l'estimateur du maximum de vraisemblance est bien défini, et c'est  $T_n = h_n(X_1, \dots, X_n) = \frac{n}{\sum_{k=1}^n X_k} = \frac{1}{\bar{X}_n}$ . On retrouve la même que l'estimateur par la méthode des moments.

Même principe dans le cas discret. Soit  $\mathcal{X}$  dénombrable. Supposons que, pour tout  $\theta \in \Theta$ , si  $X \sim \mathbb{P}_\theta$  alors  $X$  est à valeurs dans  $\mathcal{X}$ . On définit  $p : \Theta \times \mathcal{X} \rightarrow [0; 1]$  par  $p(\theta, x) = \mathbb{P}_\theta(X = x)$  pour tout  $(\theta, x) \in \Theta \times \mathcal{X}$ . Ainsi,  $(p(\theta, x))_{x \in \mathcal{X}}$  est la fonction de probabilité de  $X$ . La fonction de probabilité d'un  $n$ -échantillon  $(X_1, \dots, X_n)$  de loi-mère  $\mathbb{P}_\theta$  est alors  $(\prod_{k=1}^n p(\theta, x_k))_{(x_1, \dots, x_n) \in \mathcal{X}^n}$ .

**Définition 3.2.26** (Fonction de vraisemblance, cas discret). On définit  $p_n : \Theta \times \mathcal{X}^n \rightarrow [0; 1]$  par  $p_n : (\theta, x_1, \dots, x_n) \mapsto \prod_{k=1}^n p(\theta, x_k)$ . On appelle *fonction de vraisemblance* en  $(x_1, \dots, x_n) \in \mathcal{X}^n$  la fonction  $V_{(x_1, \dots, x_n)} : \theta \mapsto p_n(\theta, x_1, \dots, x_n)$  de  $\Theta$  dans  $[0; 1]$ .

Comme précédemment, on note  $D_n \subset \mathcal{X}^n$  l'ensemble des points  $(x_1, \dots, x_n) \in \mathcal{X}^n$  tels que  $V_{(x_1, \dots, x_n)}$  atteint son maximum en un unique point.

**Définition 3.2.27** (Estimateur du maximum de vraisemblance, cas discret). Soit  $h_n : D_n \rightarrow \Theta$  la fonction  $h_n : (x_1, \dots, x_n) \mapsto \arg \max_{\theta \in \Theta} V_{(x_1, \dots, x_n)}(\theta)$ . Si l'échantillon  $(X_1, \dots, X_n)$  est presque sûrement à valeurs dans  $D_n$ , l'*estimateur du maximum de vraisemblance* de  $\theta$  est  $T_n = h_n(X_1, \dots, X_n)$ .

*Remarque 3.2.28.* Dans la méthode du maximum de vraisemblance, il faut savoir déterminer si une fonction  $V : \Theta \rightarrow \mathbb{R}_+$  admet un unique maximum, puis déterminer où ce maximum est atteint.

- Supposons que  $\Theta \subset \mathbb{R}$  est un intervalle et que  $V$  est dérivable. Si  $V$  atteint un maximum en  $\theta \in \Theta$  alors  $V'(\theta) = 0$  ou  $\theta$  est une borne de l'intervalle. Un éventuel maximum global est donc à chercher parmi les bornes et les points critiques (i.e. les solutions de l'équation  $V'(\theta) = 0$ ). On rappelle que les points critiques ne sont pas tous des maxima globaux, il y a aussi des maxima locaux, des minima locaux et des points d'inflexion.
- Supposons que  $\Theta \subset \mathbb{R}^d$  est un ouvert et que  $V$  est différentiable. Si  $V$  atteint un maximum en  $\theta = (\theta_1, \dots, \theta_d) \in \Theta$  alors  $\theta$  est un point critique de  $V$  i.e.,  $\forall k \in \llbracket 1, d \rrbracket, \partial_k V(\theta_1, \dots, \theta_d) = 0$ . On trouve donc les candidats à être des maxima globaux en résolvant un système de  $d$  équations à  $d$  inconnues. Là encore, tous les points critiques ne sont pas des maxima globaux.

**Définition 3.2.29** (Support d'une fonction). Soit  $V : \Theta \rightarrow \mathbb{R}_+$  une fonction, on appelle *support* de  $V$  l'ensemble  $\text{Supp}(V) = \{\theta \in \Theta \mid V(\theta) > 0\}$ .

Dans de nombreux cas, la fonction de vraisemblance est un produit de fonctions puissance et exponentielle. Il peut alors être plus simple de maximiser la fonction de log-vraisemblance :  $\ln \circ V_{(x_1, \dots, x_n)}$ .

Plus précisément, soit  $(x_1, \dots, x_n) \in \mathcal{X}^n$ . Comme  $V_{(x_1, \dots, x_n)}$  est à valeurs positives ou nulles, si elle n'est pas constante à 0, alors  $\text{Supp} V_{(x_1, \dots, x_n)} \neq \emptyset$  et un éventuel maximum de  $V_{(x_1, \dots, x_n)}$  est nécessairement atteint sur ce support. Sur  $\text{Supp} V_{(x_1, \dots, x_n)}$ , la fonction  $\ln \circ V_{(x_1, \dots, x_n)}$  est bien définie. Comme  $\ln$  est strictement croissante,  $\ln \circ V_{(x_1, \dots, x_n)}$  atteint un maximum en  $\theta$  si et seulement si  $V_{(x_1, \dots, x_n)}$  atteint un maximum en  $\theta$ . Ainsi,  $(x_1, \dots, x_n) \in D_n$  si et seulement si  $\text{Supp}(V_{(x_1, \dots, x_n)}) \neq \emptyset$  et  $\ln \circ V_{(x_1, \dots, x_n)}$  atteint un unique maximum sur cet ensemble. De plus, dans ce cas,

$$h_n(x_1, \dots, x_n) = \arg \max_{\theta \in \Theta} V_{(x_1, \dots, x_n)}(\theta) = \arg \max_{\theta \in \text{Supp}(V_{(x_1, \dots, x_n)})} \ln(V_{(x_1, \dots, x_n)}(\theta)).$$

*Exemple 3.2.30.* Si on revient au modèle  $(\mathcal{E}(\theta))_{\theta > 0}$ . Soit  $(x_1, \dots, x_n) \in \mathcal{X}^n$ , où  $\mathcal{X} = ]0; +\infty[$ . On a  $\text{Supp}(V_{(x_1, \dots, x_n)}) = \Theta = ]0; +\infty[$  et, pour tout  $\theta > 0$ ,  $\ln \circ V_{(x_1, \dots, x_n)}(\theta) = n \ln(\theta) - \theta \sum_{k=1}^n x_k$ . Cette fonction  $\mathcal{C}^\infty$  tend vers  $-\infty$  en 0 et en  $+\infty$ , donc elle atteint un maximum. Puis

$$(\ln \circ V_{(x_1, \dots, x_n)})'(\theta) = 0 \iff \frac{n}{\theta} - \sum_{k=1}^n x_k = 0 \iff \theta = \frac{n}{\sum_{k=1}^n x_k}.$$

Donc le maximum est unique,  $(x_1, \dots, x_n) \in D_n$  et  $h_n(x_1, \dots, x_n) = \frac{n}{\sum_{k=1}^n x_k}$ .

### 3.3 Intervalles de confiance

On se place toujours dans un modèle statistique paramétrique  $(\mathcal{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ . On supposera ici que le paramètre est de dimension 1, c'est-à-dire  $\Theta \subset \mathbb{R}$ . Dans la section précédente, on a cherché des estimateurs du paramètre  $\theta$ , c'est-à-dire des statistiques dont les réalisations sont proches de  $\theta$ . On va maintenant s'intéresser à construire des intervalles de confiance, c'est-à-dire des intervalles aléatoires qui vont contenir  $\theta$  avec grande probabilité.

#### 3.3.1 Intervalles de confiance non-asymptotiques

**Définition 3.3.1** (Intervalle de confiance). Soit  $\gamma \in ]0; 1[$ , on appelle *intervalle de confiance (non-asymptotique)* de niveau  $\gamma$ , pour le paramètre  $\theta$ , tout intervalle aléatoire de la forme  $[T^-; T^+]$ , où  $T^-$  et  $T^+$  sont des statistiques telles que :

$$\forall \theta \in \Theta, \quad \mathbb{P}_\theta(T^- \leq \theta \leq T^+) \geq \gamma.$$

Si  $T^-$  et  $T^+$  sont à valeurs réelles, on dit que l'intervalle de confiance est *bilatère*. On autorise aussi le cas où  $T^-$  (resp.  $T^+$ ) est constante à  $-\infty$  (resp.  $+\infty$ ). On considère alors plutôt  $] - \infty; T^+]$  (resp.  $[T^-; +\infty[$ ) et on parle d'intervalle de confiance *unilatère*.

*Remarque 3.3.2.* • Les statistiques  $T^-$  et  $T^+$  sont basées sur un certain échantillon  $(X_1, \dots, X_n)$  de loi-mère  $\mathbb{P}_\theta$ . En particulier, leur loi dépend de  $\theta$ .

- Si  $S^-$  et  $S^+$  sont des statistiques tels que  $S^- \leq T^- \leq T^+ \leq S^+$  et  $[T^-; T^+]$  est un intervalle de confiance de niveau  $\gamma$  pour  $\theta$ , alors  $[S^-; S^+]$  aussi, et bien sûr  $] - \infty; +\infty[$  aussi. Pour que l'intervalle de confiance soit pertinent, il faut que sa largeur soit aussi petite que possible.
- Si on augmente le niveau  $\gamma$ , il faut en général élargir l'intervalle. Des choix classiques sont  $\gamma = 0,95$  ou  $\gamma = 0,99$ .

Souvent, les intervalles de confiance sont construits à partir d'un estimateur ponctuel de  $\theta$ . Une première méthode pour cela est d'utiliser des inégalités de concentration.

**Lemme 3.3.3.** Soient  $C > 0$  et  $T$  un estimateur  $L^2$  de  $\theta$  tel que, pour tout  $\theta \in \Theta$ ,  $R_T(\theta) \leq C$ . Pour tout  $\gamma \in ]0; 1[$ ,  $\left[ T - \sqrt{\frac{C}{1-\gamma}}; T + \sqrt{\frac{C}{1-\gamma}} \right]$  est un intervalle de confiance de niveau  $\gamma$  pour  $\theta$ .

*Démonstration.* Soit  $\gamma \in ]0; 1[$ , pour tout  $\theta \in \Theta$ , on a :

$$\mathbb{P}_\theta \left( T - \sqrt{\frac{C}{1-\gamma}} \leq \theta \leq T + \sqrt{\frac{C}{1-\gamma}} \right) = \mathbb{P}_\theta \left( |T - \theta| \leq \sqrt{\frac{C}{1-\gamma}} \right) = 1 - \mathbb{P}_\theta \left( |T - \theta| > \sqrt{\frac{C}{1-\gamma}} \right).$$

Puis, en utilisant le lemme 3.2.8, on a  $\mathbb{P}_\theta \left( |T - \theta| > \sqrt{\frac{C}{1-\gamma}} \right) \leq \frac{1-\gamma}{C} R_T(\theta) \leq 1-\gamma$ . D'où le résultat.  $\square$

*Exemple 3.3.4.* Dans l'exemple introductif,  $(\mathcal{B}(N, \theta))_{\theta \in [0; 1]}$  avec  $N = 365$  et on a utilisé l'estimateur de  $T = \frac{1}{N} \bar{X}_{25}$ . On a vérifié que  $R_T(\theta) \leq \frac{1}{100N}$  pour tout  $\theta \in [0; 1]$ , ce dont on a déduit que,

$$\mathbb{P}_\theta \left( T - \frac{1}{\sqrt{10N}} \leq \theta \leq T + \frac{1}{\sqrt{10N}} \right) \geq 0,9,$$

c'est-à-dire  $\left[ T - \frac{1}{\sqrt{10N}}; T + \frac{1}{\sqrt{10N}} \right]$  est un intervalle de confiance de niveau 0,9.

### 3.3.2 Méthode de la fonction pivot

**Définition 3.3.5** (Quantile). Soient  $X$  une v.a.r. de fonction de répartition  $F$  et  $\alpha \in ]0; 1[$ , on appelle *quantile d'ordre  $\alpha$*  de  $X$  (ou de sa loi) tout  $q_\alpha \in \mathbb{R}$  tel que  $\mathbb{P}(X \leq q_\alpha) = F(q_\alpha) = \alpha$ .

Il n'y a pas toujours existence (si  $F$  n'est pas surjective) ni unicité (si  $F$  n'est pas injective) d'un quantile d'ordre  $\alpha$  pour une loi donnée. Typiquement, ça se passe mal pour les lois discrètes. En revanche, si  $X$  admet une densité  $f$  et  $\text{Supp}(f)$  est un intervalle  $]a; b[$ , alors  $F$  est continue, nulle sur  $] - \infty; a[$ , strictement croissante sur  $]a; b[$  et constante à 1 sur  $[b; +\infty[$ . Donc  $F$  est une bijection de  $]a; b[$  vers  $]0; 1[$  et, pour tout  $\alpha \in ]0; 1[$ , il existe un unique  $q_\alpha \in ]a; b[$  tel que  $F(q_\alpha) = \alpha$ . Dans la suite, quand on parle de quantiles, on se place implicitement dans ce cadre.

*Exemple 3.3.6.* Un cas utile est  $X \sim \mathcal{N}(0, 1)$ , où  $F : x \mapsto \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt$  réalise une bijection de  $\mathbb{R}$  vers  $]0; 1[$ . Pour tout  $\alpha \in ]0; 1[$ , le quantile de niveau  $\alpha$  de la  $\mathcal{N}(0, 1)$  est uniquement défini et vaut  $q_\alpha = F^{-1}(\alpha)$ . Par parité de la densité, on a  $q_{\frac{1}{2}} = 0$  et, pour tout  $\alpha \in ]0; 1[$ ,  $q_{1-\alpha} = -q_\alpha$ , en effet,

$$F(q_{1-\alpha}) = 1 - \alpha = 1 - F(q_\alpha) = 1 - \int_{-\infty}^{q_\alpha} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} dt = \int_{q_\alpha}^{+\infty} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} dt = \int_{-\infty}^{-q_\alpha} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} dt = F(-q_\alpha).$$

Ni  $F$  ni  $F^{-1}$  n'ont d'expression analytique simple. Pour déterminer  $q_\alpha$ , on utilise des valeurs tabulées. Par exemple,  $q_{0,975} = -q_{0,025} \simeq 1,960$  et  $q_{0,995} = -q_{0,005} \simeq 2,576$ .

**Lemme 3.3.7.** *Soit  $X$  une v.a.r. dont la fonction de répartition  $F$  est continue. Soient  $q_\alpha$  et  $q_\beta$  des quantiles de  $X$  d'ordre  $\alpha$  et  $\beta$  respectivement, où  $0 < \alpha < \beta < 1$ . Alors  $\mathbb{P}(q_\alpha \leq X \leq q_\beta) = \beta - \alpha$ . En particulier, pour tout  $\gamma \in ]0; 1[$ ,  $\mathbb{P}\left(q_{\frac{1-\gamma}{2}} \leq X \leq q_{\frac{1+\gamma}{2}}\right) = \gamma$ .*

*Démonstration.* Comme  $\{\omega \in \Omega \mid q_\alpha \leq X(\omega) \leq q_\beta\} = \{\omega \in \Omega \mid X(\omega) \leq q_\beta\} \setminus \{\omega \in \Omega \mid X(\omega) < q_\alpha\}$ ,  

$$\mathbb{P}(q_\alpha \leq X \leq q_\beta) = \mathbb{P}(X \leq q_\beta) - \mathbb{P}(X < q_\alpha) = F(q_\beta) - \lim_{x \rightarrow q_\alpha^-} F(x) = F(q_\beta) - F(q_\alpha) = \beta - \alpha. \quad \square$$

*Exemple 3.3.8.* Si  $X \sim \mathcal{N}(0, 1)$ , alors  $\mathbb{P}(X \in [-1,97; 1,97]) \geq \mathbb{P}(X \in [q_{0,025}; q_{0,975}]) = 0,95$  et  $\mathbb{P}(X \in [-2,58; 2,58]) \geq \mathbb{P}(X \in [q_{0,005}; q_{0,995}]) = 0,99$ .

**Définition 3.3.9** (Fonction pivot). Soient  $n \in \mathbb{N}^*$  et  $(X_1, \dots, X_n)$  un  $n$ -échantillon de loi-mère  $\mathbb{P}_\theta$ . Une fonction  $g : \Theta \times \mathcal{X}^n \rightarrow \mathbb{R}$  est appelée *fonction pivot* si :

- la loi de  $g(\theta, X_1, \dots, X_n)$  ne dépend pas de  $\theta$ ;
- pour tout  $a < b$ , il existe des fonctions  $h^-$  et  $h^+$  de  $\mathcal{X}^n$  dans  $\Theta$  telles que :

$$\forall (x_1, \dots, x_n) \in \mathcal{X}^n, \quad a \leq g(\theta, x_1, \dots, x_n) \leq b \iff h^-(x_1, \dots, x_n) \leq \theta \leq h^+(x_1, \dots, x_n).$$

Supposons qu'une telle fonction pivot existe pour notre modèle, et notons  $Y$  une v.a.r. de même loi que la loi commune des  $(g(\theta, X_1, \dots, X_n))_{\theta \in \Theta}$ . Soient  $(q_\alpha)_{0 < \alpha < 1}$  les quantiles de la loi de  $Y$ , que l'on suppose bien définis. Soit  $\gamma \in ]0; 1[$ , pour tout  $\theta \in \Theta$ , on a

$$\mathbb{P}_\theta\left(q_{\frac{1-\gamma}{2}} \leq g(\theta, X_1, \dots, X_n) \leq q_{\frac{1+\gamma}{2}}\right) = \mathbb{P}\left(q_{\frac{1-\gamma}{2}} \leq Y \leq q_{\frac{1+\gamma}{2}}\right) = \gamma.$$

On peut ensuite pivoter en utilisant les fonctions  $h^-$  et  $h^+$  telles que

$$\forall (x_1, \dots, x_n) \in \mathcal{X}^n, \quad q_{\frac{1-\gamma}{2}} \leq g(\theta, x_1, \dots, x_n) \leq q_{\frac{1+\gamma}{2}} \iff h^-(x_1, \dots, x_n) \leq \theta \leq h^+(x_1, \dots, x_n).$$

En définissant les statistiques  $T^- = h^-(X_1, \dots, X_n)$  et  $T^+ = h^+(X_1, \dots, X_n)$ , on obtient que :

$$\forall \theta \in \Theta, \quad \mathbb{P}_\theta(T^- \leq \theta \leq T^+) = \mathbb{P}_\theta\left(q_{\frac{1-\gamma}{2}} \leq g(\theta, X_1, \dots, X_n) \leq q_{\frac{1+\gamma}{2}}\right) = \gamma,$$

c'est-à-dire  $[T^-; T^+]$  est un intervalle de confiance pour  $\theta$  de niveau  $\gamma$ .

*Exemple 3.3.10.* Considérons le modèle  $(\mathcal{N}(\theta, 1))_{\theta \in \mathbb{R}}$ , on a donc  $\Theta = \mathbb{R} = \mathcal{X}$ . Soient  $n \in \mathbb{N}^*$  et  $(X_1, \dots, X_n)$  un  $n$ -échantillon de loi-mère  $\mathcal{N}(\theta, 1)$ . On sait que  $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$  est de loi  $\mathcal{N}(\theta, \frac{1}{n})$ , donc  $\sqrt{n}(\bar{X}_n - \theta) \sim \mathcal{N}(0, 1)$ .

Posons  $g : (\theta, x_1, \dots, x_n) \mapsto \sqrt{n}(\frac{1}{n} \sum_{k=1}^n x_k - \theta)$ . Pour tout  $\theta \in \mathbb{R}$ ,  $g(\theta, X_1, \dots, X_n) \sim \mathcal{N}(0, 1)$ . De plus, soient  $(x_1, \dots, x_n) \in \mathbb{R}^n$  et  $\bar{x}_n = \frac{1}{n} \sum_{k=1}^n x_k$ , pour tout  $a < b$  on a :

$$\begin{aligned} a \leq g(\theta, x_1, \dots, x_n) \leq b &\iff \frac{a}{\sqrt{n}} \leq \bar{x}_n - \theta \leq \frac{b}{\sqrt{n}} \iff -\bar{x}_n + \frac{a}{\sqrt{n}} \leq -\theta \leq -\bar{x}_n + \frac{b}{\sqrt{n}} \\ &\iff \bar{x}_n - \frac{b}{\sqrt{n}} \leq \theta \leq \bar{x}_n - \frac{a}{\sqrt{n}}. \end{aligned}$$

Donc  $g$  est bien une fonction pivot, et les fonctions associées à  $a < b$  sont

$$h^- : (x_1, \dots, x_n) \mapsto \frac{1}{n} \sum_{k=1}^n x_k - \frac{b}{\sqrt{n}} \quad \text{et} \quad h^+ : (x_1, \dots, x_n) \mapsto \frac{1}{n} \sum_{k=1}^n x_k - \frac{a}{\sqrt{n}}.$$

Soit  $\gamma = 0,95$ , on utilise ce qui précède avec  $a = q_{\frac{1-\gamma}{2}} = q_{0,025}$  et  $b = q_{\frac{1+\gamma}{2}} = q_{0,975}$  les quantiles de la  $\mathcal{N}(0, 1)$ , on obtient  $T^- = h^-(X_1, \dots, X_n) = \bar{X}_n - \frac{q_{0,975}}{\sqrt{n}} \simeq \bar{X}_n - \frac{1,96}{\sqrt{n}}$  et  $T^+ = \bar{X}_n - \frac{q_{0,025}}{\sqrt{n}} \simeq \bar{X}_n + \frac{1,96}{\sqrt{n}}$ , et  $[T^-; T^+]$  est un intervalle de confiance de niveau 0,95 pour  $\theta$ .

### 3.3.3 Retour sur la convergence en loi

Contrairement aux autres formes de convergence, la convergence en loi se comporte mal avec les opérations : si  $X_n \xrightarrow[n \rightarrow +\infty]{\text{loi}} X$  et  $Y_n \xrightarrow[n \rightarrow +\infty]{\text{loi}} Y$ , il n'est pas vrai en général que  $X_n + Y_n \xrightarrow[n \rightarrow +\infty]{\text{loi}} X + Y$ . Par exemple, soient  $X$  et  $Y$  deux  $\mathcal{N}(0, 1)$  indépendantes, de sorte que  $X + Y \sim \mathcal{N}(0, 2)$ . Si on pose, pour tout  $n \in \mathbb{N}$ ,  $X_n = X = -Y_n$ , alors  $X_n \xrightarrow[n \rightarrow +\infty]{\text{loi}} X$ . Par ailleurs, pour tout  $n \in \mathbb{N}$ ,  $Y_n \sim \mathcal{N}(0, 1)$ , et donc  $Y_n \xrightarrow[n \rightarrow +\infty]{\text{loi}} Y$ . Cependant, pour tout  $n \in \mathbb{N}$ ,  $X_n + Y_n = 0$  et cette somme ne converge pas en loi vers  $X + Y$ . Ce genre de problème vient du fait que, pour connaître la loi de  $X + Y$ , il ne suffit pas de connaître la loi de  $X$  et la loi de  $Y$ , mais il faut connaître la loi du couple  $(X, Y)$ .

**Lemme 3.3.11.** *Soit  $(X_n)_{n \in \mathbb{N}}$  une suite de variables aléatoires réelles, qui converge en loi vers une constante  $c \in \mathbb{R}$ . Alors  $X_n \xrightarrow[n \rightarrow +\infty]{\mathbb{P}} c$ .*

*Démonstration.* Soit  $\varepsilon > 0$ , le bord de l'intervalle  $[c - \varepsilon; c + \varepsilon]$  est la paire  $\{c - \varepsilon; c + \varepsilon\}$ . Comme  $X_n \xrightarrow[n \rightarrow +\infty]{\text{loi}} c$  et  $\mathbb{P}(c \in \{c - \varepsilon; c + \varepsilon\}) = 0$ , le théorème de Portemanteau assure que

$$\mathbb{P}(X_n \in [c - \varepsilon; c + \varepsilon]) \xrightarrow[n \rightarrow +\infty]{} \mathbb{P}(c \in [c - \varepsilon; c + \varepsilon]) = 1.$$

En passant au complémentaire,

$$\mathbb{P}(|X_n - c| > \varepsilon) = \mathbb{P}(X_n \notin [c - \varepsilon; c + \varepsilon]) = 1 - \mathbb{P}(X_n \in [c - \varepsilon; c + \varepsilon]) \xrightarrow[n \rightarrow +\infty]{} 0. \quad \square$$

**Théorème 3.3.12** (Slutsky). *Soient  $(X_n)_{n \in \mathbb{N}}$  et  $(Y_n)_{n \in \mathbb{N}}$  deux suites de variables aléatoires réelles, soient  $c \in \mathbb{R}$  et  $Y$  une v.a. réelle. Si  $X_n \xrightarrow[n \rightarrow +\infty]{\text{loi}} c$  et  $Y_n \xrightarrow[n \rightarrow +\infty]{\text{loi}} Y$ , alors  $(X_n, Y_n) \xrightarrow[n \rightarrow +\infty]{\text{loi}} (c, Y)$ .*

*Démonstration.* La convergence en loi est équivalente à la convergence simple des fonctions caractéristiques. Il suffit donc de montrer que, pour tout  $s$  et  $t \in \mathbb{R}$ , on a  $\mathbb{E}[e^{i(sX_n + tY_n)}] \xrightarrow[n \rightarrow +\infty]{} \mathbb{E}[e^{i(sc + tY)}]$ . Soient  $s$  et  $t \in \mathbb{R}$ ,

$$\left| \mathbb{E}[e^{i(sX_n + tY_n)}] - \mathbb{E}[e^{i(sc + tY)}] \right| \leq \left| \mathbb{E}[e^{i(sX_n + tY_n)}] - \mathbb{E}[e^{i(sc + tY_n)}] \right| + \left| \mathbb{E}[e^{i(sc + tY_n)}] - \mathbb{E}[e^{i(sc + tY)}] \right|.$$

Pour le second terme, comme  $Y_n \xrightarrow[n \rightarrow +\infty]{\text{loi}} Y$ ,

$$\left| \mathbb{E}[e^{i(sc + tY_n)}] - \mathbb{E}[e^{i(sc + tY)}] \right| = |e^{isc} (\mathbb{E}[e^{itY_n}] - \mathbb{E}[e^{itY}])| = |\mathbb{E}[e^{itY_n}] - \mathbb{E}[e^{itY}]| \xrightarrow[n \rightarrow +\infty]{} 0.$$

Pour le premier terme,

$$\begin{aligned} \left| \mathbb{E}[e^{i(sX_n + tY_n)}] - \mathbb{E}[e^{i(sc + tY_n)}] \right| &= \left| \mathbb{E}[e^{i(sX_n + tY_n)} - e^{i(sc + tY_n)}] \right| = \left| \mathbb{E}[(e^{is(X_n - c)} - 1)e^{i(sc + tY_n)}] \right| \\ &\leq \mathbb{E} \left[ \left| e^{is(X_n - c)} - 1 \right| \right], \end{aligned}$$

et il reste à voir que ce dernier terme tend vers 0.

Soit  $\varepsilon > 0$ , par continuité de  $x \mapsto e^{isx}$ , il existe  $\delta > 0$  tel que,  $|x| \leq \delta \implies |e^{isx} - 1| \leq \varepsilon$ . Alors,

$$\begin{aligned} \mathbb{E} \left[ \left| e^{is(X_n - c)} - 1 \right| \right] &= \mathbb{E} \left[ \left| e^{is(X_n - c)} - 1 \right| \mathbf{1}_{\{|X_n - c| \leq \delta\}} \right] + \mathbb{E} \left[ \left| e^{is(X_n - c)} - 1 \right| \mathbf{1}_{\{|X_n - c| > \delta\}} \right] \\ &\leq \varepsilon + 2\mathbb{P}(|X_n - c| > \delta). \end{aligned}$$

D'après le lemme 3.3.11, on a en fait  $X_n \xrightarrow[n \rightarrow +\infty]{\mathbb{P}} c$ . Donc  $\mathbb{P}(|X_n - c| > \delta) \xrightarrow[n \rightarrow +\infty]{} 0$  et, pour tout  $n$  assez grand,  $\mathbb{E} \left[ \left| e^{is(X_n - c)} - 1 \right| \right] \leq 2\varepsilon$ . Donc  $\mathbb{E} \left[ \left| e^{is(X_n - c)} - 1 \right| \right] \xrightarrow[n \rightarrow +\infty]{} 0$ .  $\square$

**Corollaire 3.3.13.** Si  $X_n \xrightarrow[n \rightarrow +\infty]{loi} c$  et  $Y_n \xrightarrow[n \rightarrow +\infty]{loi} Y$ , alors on a  $X_n + Y_n \xrightarrow[n \rightarrow +\infty]{loi} c + Y$  et  $X_n Y_n \xrightarrow[n \rightarrow +\infty]{loi} cY$ . De plus, si  $c \neq 0$ , alors  $\frac{Y_n}{X_n} \xrightarrow[n \rightarrow +\infty]{loi} \frac{Y}{c}$ .

*Démonstration.* On fait la preuve pour l'addition, les autres cas étant similaires. Soit  $f : \mathbb{R} \rightarrow \mathbb{R}$  continue et bornée. Comme  $(x, y) \mapsto x + y$  est continue, alors  $g : (x, y) \mapsto f(x + y)$  est continue de  $\mathbb{R}^2$  dans  $\mathbb{R}$  et bornée. D'après le théorème de Slutsky,  $(X_n, Y_n) \xrightarrow[n \rightarrow +\infty]{loi} (c, Y)$ , donc

$$\mathbb{E}[f(X_n + Y_n)] = \mathbb{E}[g(X_n, Y_n)] \xrightarrow[n \rightarrow +\infty]{} \mathbb{E}[g(c, Y)] = \mathbb{E}[f(c, Y)].$$

C'est valable pour tout  $f \in \mathcal{C}_b^0(\mathbb{R})$ , donc  $X_n + Y_n \xrightarrow[n \rightarrow +\infty]{loi} c + Y$ .  $\square$

### 3.3.4 Intervalles de confiance asymptotiques

**Définition 3.3.14** (Intervalle de confiance asymptotique). Soient  $(T_n^-)_{n \in \mathbb{N}^*}$  et  $(T_n^+)_{n \in \mathbb{N}^*}$  deux suites de statistiques dans notre modèle  $(\mathcal{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ . Soit  $\gamma \in ]0; 1[$ , on dit que  $[T_n^-; T_n^+]$  est un *intervalle de confiance asymptotique* de niveau  $\gamma$ , pour le paramètre  $\theta$ , si

$$\forall \theta \in \Theta, \quad \lim_{n \rightarrow +\infty} \mathbb{P}_\theta(T_n^- \leq \theta \leq T_n^+) \geq \gamma.$$

Comme dans le cas non-asymptotique, les intervalles de confiance asymptotiques peuvent être *bilatères* ou *unilatères*.

Comme dans le cas non-asymptotique, une façon de construire des intervalles de confiance asymptotique est de partir d'une suite d'estimateurs de  $\theta$  et de quantifier leur concentration autour de  $\theta$ . Typiquement, soit  $(T_n)_{n \in \mathbb{N}^*}$  une suite d'estimateurs de  $\theta$  qui est asymptotiquement normale. Il existe donc  $(R_n)_{n \in \mathbb{N}^*}$  une suite positive, et  $\sigma : \Theta \rightarrow \mathbb{R}_+^*$  telles que :

$$\forall \theta \in \Theta, \quad R_n \frac{T_n - \theta}{\sigma(\theta)} \xrightarrow[n \rightarrow +\infty]{loi} \mathcal{N}(0, 1).$$

Notons  $(q_\alpha)_{0 < \alpha < 1}$  les quantiles de la gaussienne standard. Soit  $\theta \in \Theta$ , on a alors, pour tout  $\alpha < \beta$ ,

$$\mathbb{P}_\theta \left( q_\alpha \leq R_n \frac{T_n - \theta}{\sigma(\theta)} \leq q_\beta \right) = \mathbb{P}_\theta \left( R_n \frac{T_n - \theta}{\sigma(\theta)} \in [q_\alpha; q_\beta] \right) \xrightarrow[n \rightarrow +\infty]{} \mathbb{P}(\mathcal{N}(0, 1) \in [q_\alpha; q_\beta]) = \beta - \alpha.$$

*Remarque 3.3.15.* Dans ce raisonnement, on peut remplacer la loi  $\mathcal{N}(0, 1)$  par n'importe quelle loi limite (indépendante de  $\theta$ ) dont on connaît les quantiles.

À ce stade, on n'a pas encore un intervalle de confiance asymptotique. Si  $R_n \frac{T_n - \theta}{\sigma(\theta)}$  est de la forme  $g(\theta, X_1, \dots, X_n)$  où  $g$  est une fonction pivot et  $(X_1, \dots, X_n)$  est un  $n$ -échantillon de loi-mère  $\mathbb{P}_\theta$ , alors on peut pivoter comme dans le cas non-asymptotique.

*Exemple 3.3.16.* Si  $\sigma$  est constante à  $\sigma_0 > 0$  alors, pour tout  $n \in \mathbb{N}^*$

$$q_\alpha \leq R_n \frac{T_n - \theta}{\sigma_0} \leq q_\beta \iff T_n - q_\beta \frac{\sigma_0}{R_n} \leq \theta \leq T_n - q_\alpha \frac{\sigma_0}{R_n}.$$

On a donc  $\mathbb{P}_\theta \left( T_n - q_\beta \frac{\sigma_0}{R_n} \leq \theta \leq T_n - q_\alpha \frac{\sigma_0}{R_n} \right) \xrightarrow[n \rightarrow +\infty]{} \beta - \alpha$ , ce qui fournit un intervalle de confiance asymptotique de niveau  $\beta - \alpha$ .

En général  $\sigma$  n'est pas constante. La proposition suivante donne un énoncé pour fixer les idées, mais il faut surtout connaître la méthode preuve et savoir l'adapter.

**Proposition 3.3.17.** *Dans notre modèle statistique, soit  $(T_n)_{n \in \mathbb{N}^*}$  une suite asymptotiquement normale d'estimateur de  $\theta$ . Soient  $(R_n)_{n \in \mathbb{N}^*}$  une suite positive et  $\sigma : \Theta \rightarrow \mathbb{R}_+^*$  telles que :*

$$\forall \theta \in \Theta, \quad R_n(T_n - \theta) \xrightarrow{\text{loi}} \mathcal{N}(0, \sigma(\theta)^2).$$

On suppose que la suite  $(T_n)_{n \in \mathbb{N}^*}$  est consistante et que  $\sigma$  est continue. Alors, pour tout  $\gamma \in ]0; 1[$ ,

$$\mathbb{P}_\theta \left( T_n - q_{\frac{1+\gamma}{2}} \frac{\sigma(T_n)}{R_n} \leq \theta \leq T_n + q_{\frac{1+\gamma}{2}} \frac{\sigma(T_n)}{R_n} \right) \xrightarrow{n \rightarrow +\infty} \gamma,$$

i.e.  $\left[ T_n - q_{\frac{1+\gamma}{2}} \frac{\sigma(T_n)}{R_n}; T_n + q_{\frac{1+\gamma}{2}} \frac{\sigma(T_n)}{R_n} \right]$  est un intervalle de confiance asymptotique de niveau  $\gamma$ .

*Démonstration.* Fixons  $\theta \in \Theta$  et travaillons sous  $\mathbb{P}_\theta$ . Comme  $(T_n)_{n \in \mathbb{N}^*}$  est consistante,  $T_n \xrightarrow[n \rightarrow +\infty]{\mathbb{P}} \theta$ .

Donc, par continuité de  $\sigma$ , on a  $\sigma(T_n) \xrightarrow[n \rightarrow +\infty]{\mathbb{P}} \sigma(\theta)$ . Comme  $\sigma$  est à valeurs strictement positive, on peut appliquer le corollaire 3.3.13 du théorème de Slutsky pour le quotient :

$$R_n \frac{T_n - \theta}{\sigma(T_n)} \xrightarrow[n \rightarrow +\infty]{\text{loi}} \frac{1}{\sigma(\theta)} \mathcal{N}(0, \sigma(\theta)^2) = \mathcal{N}(0, 1).$$

Soit  $\gamma \in ]0; 1[$ , on a donc  $\mathbb{P}_\theta \left( q_{\frac{1-\gamma}{2}} \leq R_n \frac{T_n - \theta}{\sigma(T_n)} \leq q_{\frac{1+\gamma}{2}} \right) \xrightarrow{n \rightarrow +\infty} \mathbb{P} \left( q_{\frac{1-\gamma}{2}} \leq \mathcal{N}(0, 1) \leq q_{\frac{1+\gamma}{2}} \right) = \gamma$ . Comme  $q_{\frac{1-\gamma}{2}} = -q_{\frac{1+\gamma}{2}}$ , par le même calcul que précédemment,

$$q_{\frac{1-\gamma}{2}} \leq R_n \frac{T_n - \theta}{\sigma(T_n)} \leq q_{\frac{1+\gamma}{2}} \iff T_n - q_{\frac{1+\gamma}{2}} \frac{\sigma(T_n)}{R_n} \leq \theta \leq T_n + q_{\frac{1+\gamma}{2}} \frac{\sigma(T_n)}{R_n}. \quad \square$$

*Exemple 3.3.18* (Le cas des sondages). On cherche à estimer la probabilité  $\theta \in ]0; 1[$  que les individus d'une certaine population répondent oui à une question donnée. On considère donc le modèle  $(\mathcal{B}(\theta))_{0 < \theta < 1}$ . Soit  $(X_k)_{k \in \mathbb{N}^*}$  une suite de v.a.i.i.d. de loi  $\mathcal{B}(\theta)$  et pour tout  $n \in \mathbb{N}^*$ ,  $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$ .

Par la loi des grands nombres, on sait que  $\bar{X}_n \xrightarrow[n \rightarrow +\infty]{\text{p.s.}} \theta$ , donc  $(\bar{X}_n)_{n \in \mathbb{N}^*}$  est une suite fortement consistante d'estimateurs de  $\theta$ . Par ailleurs, d'après le théorème central limite, sous  $\mathbb{P}_\theta$ ,

$$\sqrt{n}(\bar{X}_n - \theta) = \frac{1}{\sqrt{n}} \sum_{k=1}^n (X_k - \theta) \xrightarrow[n \rightarrow +\infty]{\text{loi}} \mathcal{N}(0, \theta(1 - \theta)).$$

On applique la proposition 3.3.17 avec  $\gamma = 0,95$  en utilisant la valeur approchée  $q_{0,975} \simeq 1,96$ . On obtient que  $\left[ \bar{X}_n - 1,97 \sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}}; \bar{X}_n + 1,97 \sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}} \right]$  est un intervalle de confiance asymptotique de niveau 0,95 pour  $\theta$ .

La demi-largeur de cet intervalle est de l'ordre de  $2\sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}} \leq \frac{1}{\sqrt{n}}$ . Si  $t \in [0,4; 0,6]$  alors on a  $0,24 \leq t(1 - t) \leq 0,25$ . Si le paramètre que l'on cherche à estimer est de l'ordre de 0,5 alors l'inégalité précédente est quasi-optimale. Pour obtenir un intervalle de demi-largeur inférieure à 0,001, il faudrait donc  $n \geq 10^6$ . À garder en tête lorsqu'on voit des sondages dont les résultats sont donnés à 0,1% près.

Même pour  $n = 10^6$ , on n'a rien dit sur  $\mathbb{P}_\theta \left( \bar{X}_n - 1,97 \sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}} \leq \theta \leq \bar{X}_n + 1,97 \sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}} \right)$  car on n'a qu'un résultat asymptotique lorsque  $n \rightarrow +\infty$ . Pour cela il faudrait avoir une estimation de la vitesse de convergence.

## 3.4 Régression linéaire

### 3.4.1 Droite des moindres carrés

On s'intéresse au problème déterministe suivant. Soient  $n \geq 2$  et  $x_1, \dots, x_n$  des réels distincts. Soient  $y_1, \dots, y_n \in \mathbb{R}$ , on cherche la droite affine de  $\mathbb{R}^2$  la plus proche des couples  $(x_i, y_i)$ . Plus précisément, on cherche  $a$  et  $b \in \mathbb{R}$  tels que les  $ax_i + b$  sont proches des  $y_i$ . Notons  $x = (x_1, \dots, x_n)$ ,  $y = (y_1, \dots, y_n)$  et  $\mathbf{1} = (1, \dots, 1)$ . Soient  $a$  et  $b \in \mathbb{R}$ , on a

$$\|y - (ax + b\mathbf{1})\|^2 = \left\| \begin{pmatrix} y_1 - (ax_1 + b) \\ \vdots \\ y_n - (ax_n + b) \end{pmatrix} \right\|^2 = \sum_{i=1}^n (y_i - ax_i - b)^2.$$

On formalise alors notre problème en cherchant  $(a, b) \in \mathbb{R}^2$  qui minimise  $\sum_{i=1}^n (y_i - ax_i - b)^2$ , ou de façon équivalente qui minimise la distance euclidienne  $\|y - (ax + b\mathbf{1})\|$  entre  $y$  et  $ax + b\mathbf{1}$  dans  $\mathbb{R}^n$ .

Ce problème est important en sciences expérimentales. Typiquement les  $y_i$  sont les résultats de mesures d'une quantité dépendant des  $x_i$  via une fonction affine  $x \mapsto ax + b$  dont on cherche à déterminer les coefficients. Dans un cas idéal, on aurait  $y_i = ax_i + b$  pour tout  $i \in \llbracket 1, n \rrbracket$ , et on pourrait déterminer  $a$  et  $b$ . En pratique, diverses sources d'erreur expérimentales font qu'on obtient plutôt, pour tout  $i \in \llbracket 1, n \rrbracket$ ,  $y_i = ax_i + b + \eta_i$ , où  $\eta_i$  est un terme d'erreur qu'on espère petit. Notre problème revient alors à trouver  $a$  et  $b$  tel que  $\sum_{i=1}^n \eta_i^2$  soit minimale.

**Proposition 3.4.1.** *Soit  $n \geq 2$ . Soient  $y = (y_1, \dots, y_n)$  et  $x = (x_1, \dots, x_n) \in \mathbb{R}^n$  tel que  $x_1, \dots, x_n$  sont distincts, on note  $\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k$  et  $\bar{y} = \frac{1}{n} \sum_{k=1}^n y_k$ . Alors la fonction*

$$F : (a, b) \mapsto \|y - (ax + b\mathbf{1})\|^2 = \sum_{i=1}^n (y_i - ax_i - b)^2$$

atteint un unique minimum global sur  $\mathbb{R}^2$ . De plus ce minimum est atteint en unique point  $(a^*, b^*)$ , où  $a^* = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$  et  $b^* = \bar{y} - \bar{x} \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \bar{y} - \bar{x}a^*$ .

**Définition 3.4.2** (Droite des moindres carrés). La droite affine  $\{(x, a^*x + b^*) \mid x \in \mathbb{R}\} \subset \mathbb{R}^2$  est appelée *droite des moindres carrés* associée aux points  $((x_i, y_i))_{1 \leq i \leq n}$ .

*Démonstration.* Commençons par prouver l'existence d'un minimum global. Soit  $(a, b) \in \mathbb{R}^2$ , on a

$$\begin{aligned} F(a, b) &= \|y - (ax + b\mathbf{1})\|^2 = \langle y - (ax + b\mathbf{1}), y - (ax + b\mathbf{1}) \rangle = \|y\|^2 - 2\langle ax + b\mathbf{1}, y \rangle + \|ax + b\mathbf{1}\|^2 \\ &= \|y\|^2 - 2(a\langle x, y \rangle + b\langle \mathbf{1}, y \rangle) + a^2\|x\|^2 + 2ab\langle \mathbf{1}, x \rangle + b^2\|\mathbf{1}\|^2 = \|y\|^2 - 2L(a, b) + Q(a, b), \end{aligned}$$

où  $L : (a, b) \mapsto a\langle x, y \rangle + b\langle \mathbf{1}, y \rangle$  et  $Q$  est la forme quadratique sur  $\mathbb{R}^2$  associée à  $S = \begin{pmatrix} \|x\|^2 & \langle \mathbf{1}, x \rangle \\ \langle \mathbf{1}, x \rangle & \|\mathbf{1}\|^2 \end{pmatrix}$ .

La matrice  $S$  est symétrique, donc il existe  $O \in O_2(\mathbb{R})$  tel que  $S = {}^tO \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} O$ , où  $\lambda_1 \leq \lambda_2$  sont les valeurs propres de  $S$ . On a  $\det(S) = \|x\|^2\|\mathbf{1}\|^2 - \langle \mathbf{1}, x \rangle^2 \geq 0$  par l'inégalité de Cauchy-Schwarz. Le cas d'égalité correspond à  $x$  et  $\mathbf{1}$  colinéaires. S'il y avait égalité, il existerait  $c \in \mathbb{R}$  tel que  $x = c\mathbf{1}$ . Or  $x_1, \dots, x_n$  sont distincts, donc  $\det(S) = \lambda_1\lambda_2 > 0$  et  $\lambda_1$  et  $\lambda_2$  sont de même signe strict. Par ailleurs  $\text{Tr}(S) = \lambda_1 + \lambda_2 = \|x\|^2 + \|\mathbf{1}\|^2 > 0$ , donc  $0 < \lambda_1 \leq \lambda_2$ . En notant  $v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = O \begin{pmatrix} a \\ b \end{pmatrix}$ , on a :

$$\begin{aligned} Q(a, b) &= \begin{pmatrix} a & b \end{pmatrix} S \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} a & b \end{pmatrix} {}^tO \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} O \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} v_1 & v_2 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \lambda_1 v_1^2 + \lambda_2 v_2^2 \\ &\geq \lambda_1 \|v\|^2 = \lambda_1 \|O \begin{pmatrix} a \\ b \end{pmatrix}\|^2 = \lambda_1 \|(a, b)\|^2. \end{aligned}$$

En notant  $\|L\|$  la norme d'opérateur de la forme linéaire  $L$ , on a  $|L(a, b)| \leq \|L\| \|(a, b)\|$ . Finalement, pour tout  $(a, b) \in \mathbb{R}^2$ ,  $F(a, b) \geq \|y\|^2 - 2\|L\| \|(a, b)\| + \lambda_1 \|(a, b)\|^2 \xrightarrow{\|(a, b)\| \rightarrow +\infty} +\infty$  car  $\lambda_1 > 0$ . Donc

$F$  est continue sur  $\mathbb{R}^2$  et tend vers  $+\infty$  à l'infini, donc elle admet un minimum global.

Comme  $F$  est en fait  $C^\infty$ , son minimum est atteint en un point critique. Pour tout  $(a, b) \in \mathbb{R}^2$ ,

$$\partial_1 F(a, b) = -2\langle x, y \rangle + 2a\|x\|^2 + 2b\langle \mathbf{1}, x \rangle \quad \text{et} \quad \partial_2 F(a, b) = -2\langle \mathbf{1}, y \rangle + 2b\|\mathbf{1}\|^2 + 2a\langle \mathbf{1}, x \rangle,$$

et  $(a, b)$  est un point critique de  $F$  si et seulement si :

$$\begin{aligned} \begin{cases} a\|x\|^2 + b\langle \mathbf{1}, x \rangle = \langle x, y \rangle \\ a\langle \mathbf{1}, x \rangle + b\|\mathbf{1}\|^2 = \langle \mathbf{1}, y \rangle \end{cases} &\iff \begin{cases} a\|x\|^2 + bn\bar{x} = \langle x, y \rangle \\ an\bar{x} + bn = n\bar{y} \end{cases} &\iff \begin{cases} a\|x\|^2 - an\bar{x}^2 + n\bar{x}\bar{y} = \langle x, y \rangle \\ bn = n\bar{y} - an\bar{x} \end{cases} \\ &\iff \begin{cases} a(\|x\|^2 - n\bar{x}^2) = \langle x, y \rangle - n\bar{x}\bar{y} \\ b = \bar{y} - a\bar{x} \end{cases} \end{aligned}$$

On a  $\|x\|^2 - n\bar{x}^2 = \frac{1}{n}(\|x\|^2\|\mathbf{1}\|^2 - \langle \mathbf{1}, x \rangle^2) = \frac{1}{n} \det(S) \neq 0$ . Donc le système a une unique solution  $(a^*, b^*)$ , qui est l'unique point critique de  $F$ , et est nécessairement son unique minimum global. De plus,  $b^* = \bar{y} - a^*\bar{x}$  et  $a^* = \frac{\langle x, y \rangle - n\bar{x}\bar{y}}{\|x\|^2 - n\bar{x}^2}$ . Or,  $\langle x, y \rangle - n\bar{x}\bar{y} = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ , et  $\|x\|^2 - n\bar{x}^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2 = \sum_{i=1}^n (x_i - \bar{x})^2$ . Donc  $a^*$  et  $b^*$  sont comme annoncé.  $\square$

### 3.4.2 Le modèle linéaire

On considère de nouveau  $n \geq 2$  et  $x_1, \dots, x_n \in \mathbb{R}$  distincts. Pour  $i \in \llbracket 1, n \rrbracket$ , on considère les variables aléatoires  $Y_i = ax_i + b + \varepsilon_i$ , où  $a$  et  $b \in \mathbb{R}$  et  $(\varepsilon_1, \dots, \varepsilon_n)$  sont des v.a.i.i.d. que l'on suppose  $L^2$  et centrées. Notre but est désormais de construire des estimateurs de  $a$  et  $b$  basés sur  $(Y_1, \dots, Y_n)$ .

*Remarque 3.4.3.* Cette situation n'entre pas exactement dans le cadre des modèles paramétriques discutés précédemment. Notamment les  $(Y_i)_{1 \leq i \leq n}$  sont indépendants mais pas de même loi. Néanmoins, leur loi dépend uniquement des paramètres  $a$  et  $b \in \mathbb{R}$  et de la loi-mère de  $(\varepsilon_1, \dots, \varepsilon_n)$ , et les concepts utilisés pour les modèles paramétriques s'étendent naturellement.

On pense aux  $y_i$  de la section précédente comme à des réalisations des  $Y_i$ , disons  $y_i = Y_i(\omega)$ . L'erreur de mesure  $\eta_i$  est alors  $\varepsilon_i(\omega)$ . On suppose donc que les erreurs sont aléatoires, indépendantes et de même loi. Dans ce contexte, on cherche un procédé renvoyant des estimations de  $a$  et  $b$  à partir des observations  $(y_1, \dots, y_n)$ . Vue la proposition 3.4.1, on définit les estimateurs suivants.

**Définition 3.4.4** (Estimateurs des moindres carrés). Notons  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  et  $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ . On définit les estimateurs suivants, respectivement de  $a$  et  $b$  :

$$\alpha_n = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{et} \quad \beta_n = \bar{Y} - \bar{x}\alpha_n.$$

**Lemme 3.4.5.** Soit  $n \geq 2$ , alors  $\alpha_n$  (resp.  $\beta_n$ ) est un estimateur sans biais de  $a$  (resp.  $b$ ) dépendant linéairement des  $(Y_1, \dots, Y_n)$ . De plus, en notant  $\sigma^2$  la variance commune des  $(\varepsilon_i)_{1 \leq i \leq n}$ , alors

$$\text{Var}(\alpha_n) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \text{Var}(\beta_n) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{et} \quad \text{Cov}(\alpha_n, \beta_n) = -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

*Démonstration.* Déjà,  $\bar{Y}$  est combinaison linéaire des  $(Y_i)$ , donc les  $(Y_i - \bar{Y})$  aussi, et donc  $\alpha_n$  et  $\beta_n$  dépendent bien linéairement de  $(Y_1, \dots, Y_n)$ . Ensuite, fixons  $a$  et  $b \in \mathbb{R}$  et la loi commune de  $\varepsilon_1, \dots, \varepsilon_n$ , que l'on suppose de moyenne nulle et de variance  $\sigma^2$ . Ces éléments étant fixés, on a

$$\forall i \in \llbracket 1, n \rrbracket, \quad \mathbb{E}[Y_i] = \mathbb{E}[ax_i + b + \varepsilon_i] = ax_i + b + \mathbb{E}[\varepsilon_i] = ax_i + b \quad \text{et} \quad \text{Var}(Y_i) = \text{Var}(\varepsilon_i) = \sigma^2.$$

Donc  $\mathbb{E}[\bar{Y}] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_i] = \frac{1}{n} \sum_{i=1}^n (ax_i + b) = a\bar{x} + b$  et  $\mathbb{E}[Y_i - \bar{Y}] = a(x_i - \bar{x})$  pour tout  $i$ . Finalement,

$$\mathbb{E}[\alpha_n] = \frac{\sum_{i=1}^n (x_i - \bar{x}) \mathbb{E}[Y_i - \bar{Y}]}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n a(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = a \quad \text{et} \quad \mathbb{E}[\beta_n] = \mathbb{E}[\bar{Y}] - \bar{x} \mathbb{E}[\alpha_n] = b.$$

Donc  $\alpha_n$  (resp.  $\beta_n$ ) est bien un estimateur sans biais de  $a$  (resp.  $b$ ). Pour calculer les variances, on remarque déjà que  $\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - n\bar{x} = 0$  donc

$$\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^n (x_i - \bar{x})Y_i - \bar{Y} \sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n (x_i - \bar{x})Y_i.$$

Donc, comme les  $(Y_i)_{1 \leq i \leq n}$  sont indépendantes,

$$\text{Var}(\alpha_n) = \text{Var}\left(\sum_{i=1}^n \frac{(x_i - \bar{x})}{\sum_{k=1}^n (x_k - \bar{x})^2} Y_i\right) = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{(\sum_{k=1}^n (x_k - \bar{x})^2)^2} \text{Var}(Y_i) = \frac{\sigma^2}{\sum_{k=1}^n (x_k - \bar{x})^2}.$$

Les calculs de  $\text{Var}(\beta_n)$  et  $\text{Cov}(\alpha_n, \beta_n)$  sont du même type et laissés en exercice.  $\square$

**Corollaire 3.4.6** (Prédiction). *Pour tout  $t \in \mathbb{R}$ , la variable  $\alpha_n t + \beta_n$  est un estimateur de  $at + b$  basé sur les  $(Y_i)_{1 \leq i \leq n}$ , linéaire en  $(Y_1, \dots, Y_n)$ , sans biais, et de variance  $\sigma^2 \left( \frac{1}{n} + \frac{(t - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$ .*

*Démonstration.* Exercice.  $\square$

Les estimateurs des moindres carrés sont optimaux parmi les estimateurs linéaires sans biais.

**Théorème 3.4.7** (Gauss–Markov). *Soit  $S_n$  (resp.  $T_n$ ) un estimateur de  $a$  (resp.  $b$ ) basé sur les  $(Y_i)_{1 \leq i \leq n}$ , linéaire en  $(Y_1, \dots, Y_n)$  et sans biais. Pour tout  $a$  et  $b \in \mathbb{R}$ , pour tout choix de loi  $L^2$  et centrée des erreurs  $(\varepsilon_i)_{1 \leq i \leq n}$ , on a  $\text{Var}(\alpha_n) \leq \text{Var}(S_n)$  (resp.  $\text{Var}(\beta_n) \leq \text{Var}(T_n)$ ).*

*Démonstration.* Admis.  $\square$

Un cas particulier intéressant est celui où les erreurs  $(\varepsilon_i)_{1 \leq i \leq n}$  sont supposées gaussiennes, de loi commune  $\mathcal{N}(0, v)$ . On obtient alors un modèle paramétrique indexé par  $(a, b, v) \in \mathbb{R}^2 \times \mathbb{R}_+^* = \Theta$ . Pour tout  $(a, b, v) \in \Theta$  et  $n \geq 2$  les variables  $(Y_i)_{1 \leq i \leq n}$  sont alors des gaussiennes indépendantes, avec  $Y_i \sim \mathcal{N}(ax_i + b, v)$  pour tout  $i \in \llbracket 1, n \rrbracket$ . Leur loi jointe admet donc la densité suivante sur  $\mathbb{R}^n$  :

$$(y_1, \dots, y_n) \mapsto \frac{1}{(\sqrt{2\pi v})^n} \exp\left(-\sum_{i=1}^n \frac{(y_i - (ax_i + b))^2}{2v}\right).$$

Pour tout  $y = (y_1, \dots, y_n) \in \mathbb{R}^n$ , on obtient donc la fonction de log-vraisemblance :

$$\ln \circ V_y : (a, b, v) \mapsto -\frac{n}{2}(\ln(2\pi) + \ln(v)) - \frac{1}{2v} \sum_{i=1}^n (y_i - (ax_i + b))^2.$$

Des arguments similaires à ceux de la section 3.4.1 montrent que, pour tout  $(y_1, \dots, y_n) \in \mathbb{R}^n$ , cette fonction admet un maximum global en un unique point  $(a^*, b^*, v^*)$ , où  $a^* = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$ ,  $b^* = \bar{y} - \bar{x}a^*$  et  $v^* = \frac{1}{n} \sum_{i=1}^n (y_i - (a^*x_i + b^*))^2$ . On en déduit les estimateurs du maximum de vraisemblance de  $a$ ,  $b$  et  $v$ . Pour  $a$  et  $b$ , ce sont les estimateurs des moindres carrés  $\alpha_n$  et  $\beta_n$ . Pour  $v$ , c'est l'estimateur  $S_n = \frac{1}{n} \sum_{i=1}^n (Y_i - (\alpha_n x_i - \beta_n))^2$ . L'intérêt pratique de ce dernier estimateur est de fournir des estimations de l'erreur de mesure  $v = \text{Var}(\varepsilon_i)$  uniquement basées sur les observations.

# Bibliographie

- [1] M. Lejeune, *Statistique*, 2ème ed., Statistiques et Probabilités Appliquées., Springer, 2010.
- [2] J.-Y. Ouvrard, *Probabilités 1*, 2ème ed., Enseignement des mathématiques, Cassini, 2007.
- [3] ———, *Probabilités 2*, 3ème ed., Enseignement des mathématiques, Cassini, 2009.