Probabilités, Statistiques et Machine Learning

Killian Le Milbeau

22 mars 2026

Avant-propos

N'hésitez pas à m'envoyer un message en cas d'erreur : killian.le-milbeau@ens-rennes.fr

Table des matières

1 Intégration de Lebesgue

Sources : **[1]** Intégration de Lebesgue, François Bolley, ENS Rennes, 2022-2023.
[2] Real and Complex analysis, W. Rudin.

1.1 Ensembles

On considérera souvent des suites et des fonctions à valeurs dans l'ensemble $\overline{\mathbb{R}}$. On étend $\mathbb{R}$ à $\overline{\mathbb{R}}$ de manière naturelle, les opérations classiques (avec une attention particulière sur les opérations impliquant ∞) et l'ordre classique.
On définit une distance d sur $\overline{\mathbb{R}}$, en posant par exemple $d(x,y) = |f(x) - f(y)|$ pour une bijection strictement croissante $f : \overline{\mathbb{R}} \to [0,1]$.
Tout ensemble non-vide E de $\overline{\mathbb{R}}$ admet une borne supérieure et une borne inférieure. Si E est majoré (resp. minoré) alors c'est la borne supérieure (resp. inférieure) classique dans $\mathbb{R}$ sinon c'est ∞ (resp. $-\infty$).

1.1.1 Suites

Toute suite croissante (y_n) de $\overline{\mathbb{R}}$ admet une limite, égale à $\sup_n y_n$, qui est donc finie si (y_n) est bornée et ∞ sinon.

Définition 1.1.1 (Limite inférieure et supérieure)**.** Soit (x_n) une suite de $\overline{\mathbb{R}}$. La suite $(y_n) = \left(\inf_{k\geq n} x_k\right)_n$ est croissante donc admet une limite dans $\overline{\mathbb{R}}$, égale à son supremum

$$\lim_n y_n = \lim_n \inf_{k\geq n} x_k = \sup_n \inf_{k\geq n} x_k$$

notée

$$\liminf_n x_n$$

De la même manière, la suite $(z_n) = \left(\sup_{k\geq n} x_k\right)$ est décroissante donc admet une limite dans $\overline{\mathbb{R}}$, égale à son infimum

$$\lim_n z_n = \lim_n \sup_{k\geq n} x_k = \inf_n \sup_{k\geq n} x_k$$

notée

$$\limsup_n x_n$$

1.1.2 Ensembles

Soit E un ensemble, $\mathcal{P}(E)$ l'ensemble de ses parties (ie, de ses sous-ensembles) et soit $(A_n) \in \mathcal{P}(E)^{\mathbb{N}}$.

Définition 1.1.2 (Croissance et décroissance)**.** La suite (A_n) est dite croissante (au sens de l'inclusion) si $A_n \subseteq A_{n+1}$ pour tout n. On note alors sa limite

$$\lim_n A_n = \bigcup_{n=0}^{\infty} A_n$$

De la même manière, la suite (A_n) est dite décroissante (au sens de l'inclusion) si $A_{n+1} \subseteq A_n$. On note alors sa limite

$$\lim_n A_n = \bigcap_{n=0}^{\infty} A_n$$

Exemple 1.1. Dans $\mathbb{R}$, la suite $([-n, n])$ est croissante de limite $\mathbb{R}$ et la suite $([-1/n, 1/n])$ est décroissante et de limite $\{0\}$.

Définition 1.1.3 (Disjonction)**.** Des ensembles $(E_i)_{i\in I}$ sont disjoints deux à deux si

$$\forall i, j \in I, i \neq j,\ E_i \cap E_j = \emptyset$$

1.1.3 Applications

1.1.4 Dénombrabilités

1.2 Tribus

1.3 Mesures et Fonctions mesurables

1.4 Intégrales de fonctions positives

1.5 Intégrales de fonctions intégrables

1.6 Mesures produits

1.7 Changement de variables dans $\mathbb{R}^d$

1.8 Espaces L^p

1.9 Convolution dans $\mathbb{R}^d$

2 Fondements des probabilités (L3)

Sources : **[1]** Fondements des probabilités, Angst Jürgen, Université de Rennes - ENS Rennes, 2022-2023.

2.1 Espaces de probabilités

2.1.1 Espaces de probabilités

Il s'agit à l'origine de modéliser mathématiquement des phénomènes complexes dont le résultat ne peut être prédit à l'avance ou dont la modélisation déterministe est trop complexe pour être mise en oeuvre effectivement. Les exemples les plus présents dans la littérature sont les lancers de dé ou de pièce, la trajectoire d'une particule dans un gaz ou dans un liquide.

Au lieu de se focaliser sur une issue précise de l'expérience, on considère l'ensemble des résultats possibles et on leur alloue un " poids" selon qu'ils soient plus ou moins probables.

L'objet de base de la théorie est la notion d'espace de probabilités qui est la donnée d'un triplet :

$$(\Omega, \mathcal{F}, \mathbb{P})$$

L'ensemble Ω est appelé Univers. Moralement, il représente l'ensemble des résultats possibles de l'expérience qu l'on cherche à modéliser. Par exemple pour un lancer de dé, $\Omega = \{1, 2, 3, 4, 5, 6\}$. Pour la trajectoire d'une particule $\Omega = \mathcal{C}(\mathbb{R}^+, \mathbb{R}^3)$.

$\mathcal{F}$ est une tribu sur Ω, i.e. $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ et ses éléments sont appelés évènements. Moralement, les éléments de $\mathcal{F}$ sont les ensembles à qui on souhaite allouer un poids, une probabilité. Par exemple, on a la tribu triviale $\mathcal{F} = \{\emptyset, \Omega\}$, la tribu totale $\mathcal{F} = \mathcal{P}(\Omega)$. Si Ω est au plus dénombrable, on prendra $\mathcal{F} = \mathcal{P}(\Omega)$. Si Ω est un espace métrique, F =tribu borélienne.

Définition 2.1.1 (Tribu engendrée)**.** Une intersection de tribus est une tribu et si $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ alors

$$\sigma(\mathcal{A}) = \bigcap_{\mathcal{T} \text{ tribu sur } \Omega,\ \mathcal{A} \subseteq \mathcal{T}} \mathcal{T}$$

est la plus petite tribu qui contient $\mathcal{A}$. On l'appelle tribu engendrée par $\mathcal{A}$.

$\mathbb{P}$ est une mesure de probabilités sur $(\Omega, \mathcal{F})$.

Définition 2.1.2 (Mesure de probabilités)**.** On dit qu'une mesure positive $\mathbb{P}$ sur $(\Omega, \mathcal{F})$ est une mesure de probabilités si $\mathbb{P} : \mathcal{F} \to [0, 1]$ est telle que

1. $\mathbb{P}(\emptyset) = 0$
2. $\mathbb{P}(\Omega) = 1$
3. Si $(A_n) \in \mathcal{F}^{\mathbb{N}}$ avec $A_n \cap A_m = 0$ si $n \neq m$, alors

$$\mathbb{P}(\bigcup_{n \in \mathbb{N}} A_n) = \sum_{n \in \mathbb{N}} \mathbb{P}(A_n)$$

Exemple 2.1. $([0,1],\mathcal{B}([0,1]),\lambda)$; $(\Omega,\mathcal{F},\delta_a)$ où $a\in\Omega$; $(\mathbb{R},\mathcal{B}(\mathbb{R}),\mu)$ avec $\mu(dx)=\dfrac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}dx$; $(\mathbb{N},\mathcal{P}(\mathbb{N}),\mu_s)$ avec $\mu_s=\Big(\displaystyle\sum_{n\geq 1}\frac{1}{n^s}\delta_n\Big)\frac{1}{\zeta(s)}$ si $s>1$...

2.1.2 Propriétés élémentaires et opérations ensemblistes

En terme probabiliste, une union s'interprète comme un "ou" ou encore comme un "Il existe" tandis qu'une intersection s'interprète comme un "et" ou comme un "Pour tout".

Proposition 2.1.1 (Propriétés élémentaires)**.** *Soit $(\Omega,\mathcal{F},\mathbb{P})$ un espace probabilisé et A, B, $(A_n)_{n\in\mathbb{N}}$ des évènements.*

1. $\mathbb{P}(A\cup B)=\mathbb{P}(A)+\mathbb{P}(B)-\mathbb{P}(A\cap B)$;
2. *Si $B\subseteq A$ alors* $\mathbb{P}(B)\leq\mathbb{P}(A)$ *et* $\mathbb{P}(A\backslash B)=\mathbb{P}(A)-\mathbb{P}(B)$;
3. $\mathbb{P}(\overline{A})=1-\mathbb{P}(A)$;
4. $\mathbb{P}\Big(\displaystyle\bigcup_{n\in\mathbb{N}}A_n\Big)\leq\sum_{n\in\mathbb{N}}\mathbb{P}(A_n)$ *(sous-additivité)* ;
5. *Si (A_n) est croissante, $\forall n$ $A_n\subseteq A_{n+1}$, alors*
$$\mathbb{P}\Big(\bigcup_{n\in\mathbb{N}}A_n\Big)=\lim_{n\to+\infty}\mathbb{P}(A_n)=\sup_{n\in\mathbb{N}}\mathbb{P}(A_n)$$
6. *Si (A_n) est décroissante, $\forall n$ $A_{n+1}\subseteq A_n$, alors*
$$\mathbb{P}\Big(\bigcap_{n\in\mathbb{N}}A_n\Big)=\lim_{n\to+\infty}\mathbb{P}(A_n)=\inf_{n\in\mathbb{N}}\mathbb{P}(A_n)$$

<u>*Démonstration :*</u>

(a) On a $A\cup B=A\backslash(A\cap B)\sqcup B\backslash(A\cap B)\sqcup(A\cap B)$,
D'où $\mathbb{P}(A\cup B)=\mathbb{P}(A\backslash(A\cap B))+\mathbb{P}(B\backslash(A\cap B))+\mathbb{P}(A\cap B)$
$A=(A\cap B)\sqcup A\backslash(A\cap B)$ et $B=(A\cap B)\sqcup B\backslash(A\cap B)$.
Ainsi, $\mathbb{P}(A)=\mathbb{P}(A\cap B)+\mathbb{P}(A\backslash(A\cap B))$ et $\mathbb{P}(B)=\mathbb{P}(A\cap B)+\mathbb{P}(B\backslash(A\cap B))$.
Finalement, $\mathbb{P}(A\cup B)=\mathbb{P}(A)+\mathbb{P}(B)-\mathbb{P}(A\cap B)$.

(b) Si $B\subseteq A$ alors $A=B\cup(A\backslash B)$.
D'où par ce qui précède, $\mathbb{P}(A)=\mathbb{P}(B)+\mathbb{P}(A\backslash B)-\mathbb{P}(B\cap(A\backslash B))=\mathbb{P}(B)+\mathbb{P}(A\backslash B)$.
Puis, comme $\mathbb{P}$ est une mesure de probabilité et donc positive, $\mathbb{P}(A\backslash B)\geq 0$.
Finalement, $\mathbb{P}(B)\leq\mathbb{P}(A)$.
Pour le deuxième point, on reprend
$\mathbb{P}(A)=\mathbb{P}(B)+\mathbb{P}(A\backslash B)-\mathbb{P}(B\cap(A\backslash B))=\mathbb{P}(B)+\mathbb{P}(A\backslash B)$.

(c) On sait que $\Omega=A\sqcup\overline{A}$ et $\mathbb{P}(\Omega)=1$.

(d) Par le premier point, $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.
Par récurrence immédiate, pour tout n on a $\mathbb{P}(A_1 \cup ... \cup A_n) \leq \mathbb{P}(A_1) + ... + \mathbb{P}(A_n)$.
Pour passer à la limite, on a besoin du point suivant et le résultat tombe tout seul.

(e) Si (A_n) est une suite croissante, on pose comme convention $A_{-1} = \emptyset$.

$$A_n = \bigcup_{k=0}^{n} A_k = \bigsqcup_{k=0}^{n} A_k \backslash A_{k-1}$$

$$\bigcup_{k=0}^{\infty} A_k = \bigsqcup_{k=0}^{\infty} A_k \backslash A_{k-1}$$

Si bien que,

$$\mathbb{P}(A_n) = \sum_{k=0}^{n} \mathbb{P}(A_k \backslash A_{k-1})$$

Et, par σ-additivité :

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \sum_{k=0}^{\infty} \mathbb{P}(A_k \backslash A_{k-1}) = \mathbb{P}\big(\bigsqcup_{k=0}^{\infty} A_k \backslash A_{k-1}\big) = \mathbb{P}\big(\bigcup_{n=0}^{\infty} A_n\big)$$

(f) On conclut en passant au complémentaire dans le point précédent.□

On rappelle que dans le cadre d'une suite réelle (x_n),

$$\liminf_{n\to\infty} x_n = \sup_{n\geq 0} \inf_{k\geq n} x_k$$

$$\limsup_{n\to\infty} x_n = \inf_{n\geq 0} \sup_{k\geq n} x_k$$

Définition 2.1.3 (Limites inférieures et supérieures)**.** Si (A_n) est une suite d'évènements, on pose

$$\liminf_{n\to\infty} A_n = \bigcup_{n\geq 1} \bigcap_{k\geq n} A_k$$

$$\limsup_{n\to\infty} A_n = \bigcap_{n\geq 1} \bigcup_{k\geq n} A_k$$

Premièrement, on note que les limites inférieures et supérieures d'une suite d'ensemble est un ensemble.

Puis on note que moralement,

$$\omega \in \liminf A_n \Leftrightarrow \exists n \geq 1,\ \forall k \geq n,\ w \in A_k \Leftrightarrow \omega \text{ est dans tous les } A_k \text{ APCR}$$

$$\omega \in \limsup A_n \Leftrightarrow \forall n \geq 1,\ \exists k \geq n,\ \omega \in A_k \Leftrightarrow \omega \text{ est dans une infinité de } A_k$$

De plus,

$$\overline{\liminf A_n} = \overline{\bigcup_{n\geq 1}\bigcap_{k\geq n} A_k} = \bigcap_{n\geq 1}\bigcup_{k\geq n} A_k = \limsup A_n$$

Proposition 2.1.2. *Soit (A_n) une suite d'évènements, on a*

$$\liminf_{n\to\infty} \mathbb{1}_{A_n} = \mathbb{1}_{\liminf A_n}$$

$$\limsup_{n\to\infty} \mathbb{1}_{A_n} = \mathbb{1}_{\limsup A_n}$$

Démonstration : Par passage au complémentaire, on peut se restreindre à la limsup.
De plus, on remarque que $\limsup_{n\to\infty} \mathbb{1}_{A_n}$ est une limsup de suite et que $\mathbb{1}_{\limsup A_n}$ est indexé par une limsup d'ensemble. Nous allons donc devoir jouer sur les deux définitions avec précaution.
On a les équivalences suivantes :

$$\omega \in \limsup A_n \Leftrightarrow \forall n \geq 0,\ \exists k \geq n,\ \omega \in A_k \Leftrightarrow \forall n,\ \exists k \geq n,\ \mathbb{1}_{A_k}(\omega) = 1$$

D'où,

$$\omega \in \limsup A_n \Leftrightarrow \inf_{n\geq 0}\sup_{k\geq n} \mathbb{1}_{A_k}(\omega) = 1 \Leftrightarrow \limsup \mathbb{1}_{A_k}(\omega) = 1$$

□

Définition 2.1.4 (Suite d'évènements convergente)**.** Une suite d'évènements (A_n) est dite convergente si $\liminf A_n = \limsup A_n$, auquel cas la limite commune est notée $\lim A_n$.

Proposition 2.1.3 (Inégalité de limites)**.** *Soit (A_n) une suite d'évènements, alors*

$$\mathbb{P}(\liminf A_n) \leq \liminf \mathbb{P}(A_n) \leq \limsup \mathbb{P}(A_n) \leq \mathbb{P}(\limsup A_n)$$

En particulier, si (A_n) est convergente on a

$$\mathbb{P}(\lim A_n) = \lim \mathbb{P}(A_n)$$

Démonstration : On revient aux définitions

$$\liminf_n A_n = \bigcup_{n\geq 0}\bigcap_{k\geq n} A_k$$

On pose $C_n = \bigcap_{k\geq n} A_k$, croissante en n. D'apès ce qui précède,

$$\mathbb{P}(\liminf_n A_n) = \lim_n \mathbb{P}(C_n)$$

Or,

$$\forall p \geq 0,\ \mathbb{P}(C_n) \leq \mathbb{P}(A_{n+p})$$

$$\mathbb{P}(C_n) \leq \liminf_p \mathbb{P}(A_{n+p}) = \liminf_p \mathbb{P}(A_p)$$

Si bien que,

$$\liminf_n \mathbb{P}(C_n) \leq \liminf_p \mathbb{P}(A_p)$$

Et,

$$\lim_n \mathbb{P}(C_n) = \mathbb{P}(\lim_n A_n)$$

D'où,

$$\mathbb{P}(\liminf_n A_n) \leq \liminf_p \mathbb{P}(A_p)$$

On a donc la première inégalité.
La deuxième inégalité est évidente.
De même, on a

$$\limsup_n A_n = \bigcap_{n\geq 0} \bigcup_{k\geq n} A_k$$

On pose, $D_n = \bigcup_{k\geq n} A_k$ décroissante en n. Cette fois,

$$\forall p \geq 0,\ \mathbb{P}(D_n) \geq \mathbb{P}(A_{n+p})$$

Et,

$$\mathbb{P}(D_n) \geq \limsup_p \mathbb{P}(A_{n+p}) = \limsup_p \mathbb{P}(A_p)$$

$$\lim_n \mathbb{P}(D_n) = \mathbb{P}(\limsup_n A_n)$$

Donc,

$$\limsup_p \mathbb{P}(A_p) \leq \mathbb{P}(\limsup_n A_n)$$

On a donc la troisème inégalité.

Par définition de la limite, on a l'égalité dans le cas convergent.□

2.1.3 Complétion et prolongement des mesures - Ensembles négligeables, tribu complétée

Dans la suite, on fixe un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$.

Définition 2.1.5 (Ensembles négligeables)**.** Un ensemble $N \in \mathcal{P}(\Omega)$ est dit négligeable s'il existe $A \in \mathcal{F}$, $N \subseteq A$ et $\mathbb{P}(A) = 0$.
On dira que deux ensembles sont égaux p.s. si $A\Delta B = A\backslash B \sqcup B\backslash A$ (différence symétrique) est négligeable.

Proposition 2.1.4 (Propriétés de l'ensemble des parties négligeables)**.** *Soit $\mathcal{N} \subseteq \mathcal{P}(\Omega)$ l'ensemble des parties négligeables de Ω. Alors,*

1. $\emptyset \in \mathcal{N}$;
2. *Si* $B \subseteq A$ *avec* $A \in \mathcal{N}$, *alors* $B \in \mathcal{N}$;
3. *Si* $(A_n) \in \mathcal{N}^{\mathbb{N}}$ *alors* $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{N}$; *(Union quelconque)*
4. *Si* $(A_i) \in \mathcal{N}^I$ *alors* $\bigcap_{i\in I} A_i \in \mathcal{N}$; *(Intersection finie)*

Démonstration : (a) Evident ;

(b) Si $A \in \mathcal{N}$, alors il existe $F \in \mathcal{F}$ avec $A \subseteq F$ et tel que $\mathbb{P}(F) = 0$. Donc, il en va de même pour B puisque $B \subseteq A \subseteq F$;

(c) Si $(A_n) \in \mathcal{N}^{\mathbb{N}}$, alors pour tout $n \in \mathbb{N}$ il existe $C_n \in \mathcal{F}$, $A_n \subseteq C_n$ et $\mathbb{P}(C_n) = 0$. Ainsi, $\bigcup_{n\geq 0} A_n \subseteq \bigcup_{n\geq 0} C_n \in \mathcal{F}$ et par sous-additivité et positivité

$$0 \leq \mathbb{P}\big(\bigcup_n C_n\big) \leq \sum_n \mathbb{P}(C_n) = 0$$

(d) Si $(A_i) \in \mathcal{N}^I$ et si $i_0 \in I$, $\bigcap_i A_i \subseteq A_{i_0} \subseteq C_{i_0}$ avec $C_{i_0} \in \mathcal{F}$ et $\mathbb{P}(C_{i_0}) = 0$.□

Définition 2.1.6 (Tribu complétée)**.** On appelle tribu complétée de $\mathcal{F}$ par rapport à $\mathbb{P}$, la tribu

$$\overline{\mathcal{F}} = \sigma(\mathcal{F} \cup \mathcal{N})$$

C'est la plus petite tribu contenant A et tous les ensembles négligeables. Elle peut être construite plus explicitement de la manière suivante :

$$\overline{\mathcal{F}} = \{A \cup N \mid A \in \mathcal{F},\ N \subseteq B \text{ pour un certain } B \in \mathcal{F} \text{ avec } \mathbb{P}(B) = 0\}$$

En d'autres termes, elle contient tous les ensembles de $\mathcal{F}$ ainsi que tous les ensembles qui peuvent être obtenus en ajoutant des ensembles négligeables à des ensembles de $\mathcal{F}$.

Ainsi par cette construction on peut directement voir que $\mathcal{F} \subseteq \overline{\mathcal{F}}$ et vérifier que $\overline{\mathcal{F}}$ est bien une tribu.

Proposition 2.1.5. *On a les équivalences suivantes*

1. $A \in \overline{\mathcal{F}}$
2. $\exists B, C \in \mathcal{F},\ B \subseteq A \subseteq C,\ \mathbb{P}(C\backslash B) = 0$
3. $\exists B \in \mathcal{F},\ N \in \mathcal{N} \mid A = B \cup N$
4. $\exists B \in \mathcal{F},\ A = B$ *p.s. ie,* $A\Delta B \in \mathcal{N}$.

Démonstration : 2. $\Longrightarrow$ *3. :* Supposons qu'il existe $B, C \in \mathcal{F}$ avec $B \subseteq A \subseteq C$ avec $\mathbb{P}(C\backslash B) = 0$. Alors, $A = B \cup (A\backslash B)$ avec $B \in \mathcal{F}$ et $A\backslash B \subseteq C\backslash B$ et $\mathbb{P}(C\backslash B) = 0$. C'est-à-dire que $A\backslash B \in \mathcal{N}$, i.e. $A = B \cup \mathcal{N}$.

3. $\Longrightarrow$ *4. :* S'il existe $(B, N) \in \mathcal{F} \times \mathcal{N}$ tels que $A = B \cup N$ alors la différence symétrique s'écrit $A\Delta B = (B \cup N)\backslash B \subseteq N \in \mathcal{N}$.
C'est-à-dire que $A = B$ p.s.

4. $\Longrightarrow$ *2. :* S'il existe $B \in \mathcal{F}$ tel que $A\Delta B \in \mathcal{N}$, alors par définition il existe $D \in \mathcal{F}$ tel que $A\Delta B \subseteq D$ et $\mathbb{P}(D) = 0$.
On pose $B' = B \cap \overline{D},\ C' = B \cup D \in \mathcal{F}$.
Alors, $B' \subseteq A \subseteq C'$, $\mathbb{P}(C'\backslash B') = \mathbb{P}(D) = 0$.

3.$\Longleftrightarrow$*1. :* On va montrer que

$$T = \{\ A \in \overline{\mathcal{F}} \mid \exists (B, N) \in \mathcal{F} \times \mathcal{N},\ A = B \cup N\}$$

est une tribu qui coïncide avec $\overline{\mathcal{F}}$.
On a évidemment que $\Omega \in T$ car $\Omega = \Omega \cup \emptyset$ avec $\Omega \in \mathcal{F}$ et $\emptyset \in \mathcal{N}$.
Puis, si $A \in T$ avec $A = B \cup N$ avec $B \in \mathcal{F}$, $N \in \mathcal{N}$ c'est-à-dire qu'il existe $D \in \mathcal{F}$, $N \subseteq D$ et $\mathbb{P}(D) = 0$. Alors,

$$\overline{A} = \overline{(B \cup N)} = \overline{B} \cap \overline{N} = (\overline{B} \cap \overline{D}) \sqcup (\overline{B} \cap \overline{N} \cap D)$$

avec $\overline{B} \cap \overline{D} \in \mathcal{F}$ et $\overline{B} \cap \overline{N} \cap D \in \mathcal{N}$ i.e. $\overline{A} \in T$.
Maintenant, si $(A_n) \in T$, $A_n = B_n \cup N_n$, $B_n \in \mathcal{F}$, $N_n \in \mathcal{N}$. Alors,

$$\bigcup_{n\in\mathbb{N}} A_n = \left(\bigcup_{n\in\mathbb{N}} B_n\right) \cup \left(\bigcup_{n\in\mathbb{N}} N_n\right) \in T$$

car, $\bigcup_{n\in\mathbb{N}} B_n \in \mathcal{F}$ et $\bigcup_{n\in\mathbb{N}} N_n \in \mathcal{N}$.
Aini, T est bien une tribu et par définition on a $T \subseteq \overline{F}$.
Par ailleurs, $\mathcal{F} \subseteq T$ et $\mathcal{N} \subseteq T$ donc $\mathcal{F} \cup \mathcal{N} \subseteq T$. Ainsi, par minimalité,

$$\overline{F} = \sigma(\mathcal{F} \cup \mathcal{N}) \subseteq T$$

Si bien que, $T = \overline{\mathcal{F}}$. $\square$

Proposition 2.1.6. *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé et $\overline{F}$ la tribu complétée par rapport à $\mathbb{P}$. Alors la probabilité*

$$\begin{array}{rccc} \overline{\mathbb{P}} & : & \overline{F} & \to & [0,1] \\ & & A = B \cup N & \mapsto & \mathbb{P}(B) \end{array}$$

est bien définie et est l'unique extension de $\mathbb{P}$ à $\overline{\mathcal{F}}$ i.e. $\overline{\mathbb{P}}_{|\mathcal{F}} = \mathbb{P}$.

Démonstration : $\overline{\mathbb{P}}$ est bien définie. En effet, si $A \in \overline{\mathcal{F}}$, $A = B \cup N = B' \cup N'$ alors

$$B\Delta B' \subseteq N \cup N' \text{ i.e. } \mathbb{P}(B\Delta B') = 0$$

i.e. $\mathbb{P}(B) = \mathbb{P}(B')$.
$\overline{\mathbb{P}}$ est une probabilité. Vérifions les axiomes.

$\overline{\mathbb{P}}(\emptyset) = \mathbb{P}(\emptyset) = 0$ car $\emptyset \in \mathcal{F}$ et $\mathbb{P}$ est une probabilité.
Puis, soit $(A_n) \in \overline{\mathcal{F}}$ disjoints. $A_n = B_n \cup N_n$ et $\bigcup_{n\in\mathbb{N}} A_n = \Big(\bigcup_{n\in\mathbb{N}} B_n\Big) \cup \Big(\bigcup_{n\in\mathbb{N}} N_n\Big)$.
Avec $\bigcup_{n\in\mathbb{N}} B_n \in \mathcal{F}$ et $\bigcup_{n\in\mathbb{N}} N_n \in \mathcal{N}$.
Comme $\mathbb{P}$ est une probabilité, on utilise la σ-additivité

$$\overline{\mathbb{P}}\Big(\bigsqcup_n A_n\Big) = \mathbb{P}\Big(\bigsqcup_n B_n\Big) = \sum_n \mathbb{P}(B_n) = \sum_n \overline{\mathbb{P}}(A_n)$$

C'est-à-dire $\overline{\mathbb{P}}$ est σ-additive.
Maintenant, traitons de l'unicité. si $\overline{\mathbb{Q}}$ est une autre extension de $\mathbb{P}$ à $\overline{\mathcal{F}}$ et $A \in \overline{\mathcal{F}}$, $A = B \cup N$, $B \in \mathcal{F}$, $N \in \mathcal{N}$ i.e. il existe $D \in \mathcal{F}$, $N \subseteq D$, $\mathbb{P}(D) = 0$.
Alors $B \subseteq A \subseteq B \cup D$.

$$\overline{\mathbb{P}}(B) = \overline{\mathbb{Q}}(B) \leq \overline{\mathbb{Q}}(A) \leq \overline{\mathbb{Q}}(B \cup D) \leq \overline{\mathbb{Q}}(B) + \overline{\mathbb{Q}}(D) = \overline{\mathbb{P}}(B) + \overline{\mathbb{P}}(D) = \overline{\mathbb{P}}(B)$$

C'est-à-dire $\overline{\mathbb{Q}}(A) = \overline{\mathbb{P}}(B) = \overline{\mathbb{P}}(A)$. □

Proposition 2.1.7. *L'ensemble $\overline{\mathcal{N}}$ des négligeables pour $\overline{\mathbb{P}}$ coïncide avec $\mathcal{N}$.*

Démonstration : Par définition de $\overline{\mathbb{P}}$, on a $\mathcal{N} \subseteq \overline{\mathcal{N}}$. Traitons l'autre inclusion.
Soit $N \in \overline{\mathcal{N}}$, il existe $D \in \overline{\mathcal{F}}$, $N \subseteq D$ avec $\overline{\mathbb{P}}(D) = 0$.
i.e. $\exists B, N' \in \mathcal{F} \times \mathcal{N}$, $D = B \cup N'$ et $0 = \overline{\mathbb{P}}(D) = \mathbb{P}(B)$ donc $B \in \mathcal{N}$.
En conclusion, $N \subseteq D = B \cup N' \in \mathcal{N}$.
Si bien que $\overline{\mathcal{N}} \subseteq \mathcal{N}$. D'où le résultat. □

C'est-à-dire que la probabilité $\mathbb{P}$ peut être étendue de manière unique à $\overline{\mathbb{P}}$ en définissant $\mathbb{P}(A \cup N) = \mathbb{P}(A)$ pour $A \in \mathcal{F}$ et N négligeable.

En résumé, la tribu complétée est une extension de la tribu initiale qui inclut tous les ensembles négligeables, ce qui permet de travailler avec une mesure de probabilité qui est complète, c'est-à-dire que tous les sous-ensembles d'ensembles de mesure nulle sont également mesurables.

2.1.4 Complétion et prolongement des mesures - Lemme des classes monotones

On rappelle maintenant à présent un autre procédé d'extension classique.

Définition 2.1.7 (Classe monotone)**.** Une famille $\mathcal{M} \subseteq \mathcal{P}(\Omega)$ est une classe monotone si

1. $\Omega \in \mathcal{M}$;
2. Si $A \subseteq B$, $A, B \in \mathcal{M}$ alors $B \backslash A \in \mathcal{F}$;
3. $\mathcal{M}$ est stable par union dénombrable croissante. C'est-à-dire que si $(A_n) \in \mathcal{M}$ est croissante au sens de l'inclusion alors $\Big(\bigcup_n A_n\Big) \in \mathcal{M}$.

Une intersection de classes monotones est une classe monotone. Comme pour les tribus, on définit la classe monotone engendrée par $\mathcal{A}$

$$\mathcal{M}(\mathcal{A}) = \bigcap_{\text{classe } \mathcal{M},\ \mathcal{A}\subseteq\mathcal{M}} \mathcal{M}$$

C'est la plus petite classe monotone engendrée par $\mathcal{A}$.

Une tribu est une classe monotone par définition. De même, une classe monotone stable par intersection finie est une tribu.

Définition 2.1.8 (π-système)**.** Une classe $\mathcal{C}$ de parties d'un ensemble Ω est appelée π-système si cette classe est stable par intersection finie.

Par exemple, la classe des intervalles (demi-droites) $\mathcal{C} = \{]-\infty, x] \mid x \in \mathbb{R}\}$ est une classe monotone.

Théorème 2.1.1 (Lemme des classes monotones)**.** *Soit $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ un π-système. Alors,*

$$\mathcal{M}(\mathcal{A}) = \sigma(\mathcal{A})$$

Démonstration : Tout d'abord, $\sigma(\mathcal{A})$ est une tribu, c'est donc une classe monotone qui contient $\mathcal{A}$ par définition. Ainsi par minimalité, on a

$$\mathcal{M}(\mathcal{A}) \subseteq \sigma(\mathcal{A})$$

Pour l'autre inclusion, il suffit de montrer que $\mathcal{M}(\mathcal{A})$ est stable par intersection finie. On introduit alors

$$\mathcal{M}_1 = \{A \in \mathcal{M}(A) \mid \forall B \in \mathcal{A},\ A \cap B \in \mathcal{M}(\mathcal{A})\}$$

Montrons que $\mathcal{M}_1$ est une classe monotone.
D'abord, $\Omega \in \mathcal{M}_1$ car pour tout $B \in \mathcal{A}$, $\Omega \cap B = B \in \mathcal{A} \subseteq \mathcal{M}(\mathcal{A})$.
Puis, si $A_1, A_2 \in \mathcal{M}_1$ avec $A_1 \subseteq A_2$ si $B \in \mathcal{A}$

$$(A_2 \backslash A_1) \cap B = (A_2 \cap B) \backslash (A_1 \cap B) \in \mathcal{M}(\mathcal{A})$$

car $A_2 \cap B \in \mathcal{M}(\mathcal{A})$ et $A_1 \cap B \in \mathcal{M}(\mathcal{A})$. Donc, $A_2 \backslash A_1 \in \mathcal{M}_1$.
Finalement, si $(A_n) \in \mathcal{M}_1$ croisante, si $B \in \mathcal{A}$ alors

$$\left(\bigcup_n A_n\right) \cap B = \bigcup_n (A_n \cap B) \in \mathcal{M}(\mathcal{A})$$

car $(A_n \cap B) \in \mathcal{M}(\mathcal{A})$ croissante. C'est-à-dire, $\left(\bigcup_n A_n\right) \in \mathcal{M}_1$.
Ainsi, $\mathcal{M}_1$ est bien une classe monotone, elle contient $\mathcal{A}$ donc elle contient $\mathcal{M}(\mathcal{A})$.

Par ailleurs, par définition $\mathcal{M}_1 \subseteq \mathcal{M}(\mathcal{A})$ et donc $\mathcal{M}_1 = \mathcal{M}(\mathcal{A})$.
On pose alors,

$$\mathcal{M}_2 = \{A \in \mathcal{M}(\mathcal{A}) \mid \forall B \in \mathcal{M}(\mathcal{A}),\ A \cap B \in \mathcal{M}(\mathcal{A})\}$$

Comme ci-dessus, on montre que $\mathcal{M}_2$ est une classe monotone. Par ailleurs $\mathcal{M}_2$ contient $\mathcal{A}$. En effet, si $A \in \mathcal{A}$ et $B \in \mathcal{M}(\mathcal{A}) = \mathcal{M}_1$, $A \cap B \in \mathcal{M}(\mathcal{A})$ par définition de $\mathcal{M}_1$.
Par minimalité, on a alors $\mathcal{M}(\mathcal{A}) \subseteq \mathcal{M}_2$.
Alors comme plus haut, par définition $\mathcal{M}_2 \subseteq \mathcal{M}(\mathcal{A})$. D'où

$$\mathcal{M}_2 = \mathcal{M}(\mathcal{A})$$

Cela revient à dire que $\mathcal{M}(\mathcal{A})$ est stable par intersection, c'est donc une tribu et par minimalité

$$\sigma(\mathcal{A}) \subseteq \mathcal{M}(\mathcal{A})$$

D'où l'égalité par double inclusion. □

Le lemme des classes monotones permet de montrer, de manière économique, l'égalité entre deux lois de probabilités : de même que deux applications linéaires qui coïncident sur une base coïncident sur l'espace entier, deux mesures de probabilité qui coïncident sur un π-système coïncident sur la tribu engendrée par ce π-système.

Proposition 2.1.8 (Lemme d'unicité des mesures de probabilités)**.** *Soient $\mathbb{P}, \mathbb{Q}$ deux probabilités sur $(\Omega, \mathcal{F})$ qui coïncident sur un ensemble de partie $\mathcal{A} \subseteq \mathcal{F}$ stable par intersection. Alors, $\mathbb{P} = \mathbb{Q}$ sur $\sigma(\mathcal{A})$.*

Démonstration : On pose

$$\mathcal{M} = \{A \in \mathcal{F} \mid \mathbb{P}(A) = \mathbb{Q}(A)\}$$

On vérifie immédiatement que $\mathcal{M}$ est une classe monotone contenant $\mathcal{A}$, $\mathcal{M}$ est la plus petite classe monotone contenant $\mathcal{A}$ à savoir $\sigma(\mathcal{A})$ par le lemme des classes monotones. □

On reprend l'exemple des demi-droites. Soient μ et ν deux mesures de probabilité qui coïncident sur les demi-droites. C'est-à-dire que pour tout $x \in \mathbb{R}$,

$$\mu(]-\infty, x]) = \nu(]-\infty, x])$$

Dans ce cas, $\mu = \nu$.

Pour résumer l'utilité du lemme des classes monotones,il permet d'abord une extension de propriétés : on étend des propriétés d'une algèbre à la tribu générée par cette algèbre. Il permet aussi donc de montrer l'égalité entre deux mesures de probabilités : si deux mesures de probabilités coïncident sur une algèbre $\mathcal{A}$ et si elles sont toutes deux des mesures de probabilités sur la tribu $\sigma(\mathcal{A})$, alors elles coïncident sur $\sigma(\mathcal{A})$. Enfin, le lemme permet de simplifier les preuves en se concentrant sur une classe plus petite et plus maniable d'ensembles (l'algèbre $\mathcal{A}$) plutôt que sur la tribu entière $\sigma(\mathcal{A})$.

Théorème 2.1.2 (Version fonctionnelle du lemme des classes monotones)**.** *Soit H un espace de fonctions réelles bornées sur Ω et $\mathcal{A}$ un π-système qui contient Ω. On suppose*

- $\forall A \in \mathcal{A},\ \mathbb{1}_A \in H$;
- *Si (f_n) est une suite croissante de fonctions positives de H et si $f_n \to f$, alors $f \in H$;*

Alors, H contient toutes les fonctions $\sigma(\mathcal{A}))$-mesurables bornées.

Démonstration : On considère $\mathcal{M} = \{A \in \mathcal{P}(\Omega) \mid \mathbb{1}_A \in H\}$.
D'une part, comme $\Omega \in \mathcal{A}$, $\mathbb{1}_A \in H$ ie $\Omega \in \mathcal{M}$.
Si $A \subseteq B \in \mathcal{M}$ alors $\mathbb{1}_{B\setminus A} = \mathbb{1}_B - \mathbb{1}_A \in H$ car c'est un espace vectoriel.
Si $(A_n) \in H$ croissante, $\mathbb{1}_{\bigcup A_n} = \lim \mathbb{1}_{A_n} \in H$.
Ainsi, $\mathcal{M}$ est une classe monotone qui contient $\mathcal{A}$. De plus c'est un π-système donc $\sigma(\mathcal{A}) = \mathcal{M}(\mathcal{A}) \subseteq \mathcal{M}$.
Maintenant, soit f une fonction $\sigma(\mathcal{A})$-mesurable bornée, quitte à prendre les parties positive et négatives, on peut supposer $f \geq 0$ et $f = \lim f_n$ avec (f_n) étagée $\sigma(\mathcal{A})$-mesurable.
Alors, $(f_n) \in H$ et donc $f \in H$. D'où le résultat.

2.2 Variables aléatoires

2.2.1 Définitions et premières propriétés - Variables et vecteurs aléatoires

Définition 2.2.1 (Variable aléatoire)**.** Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé et $(E, \mathcal{E})$ un espace mesurable. On appelle variable aléatoire à valeurs dans E toute application mesurable

$$X\ :\ (\Omega, \mathcal{F}, \mathbb{P})\ \to\ (E, \mathcal{E})$$

$\forall B \in \mathcal{E},\ X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\} \in \mathcal{F}$.

Si $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, on parlera de variable aléatoire réelle.
Si $(E, \mathcal{E}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, on dira que $X = (X_1, ..., X_n)$ est un vecteur aléatoire. L'application mesurable $X_i\ :\ (\Omega, \mathcal{F}, \mathbb{P})\ \to\ (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ est appelée i-ième marginale du vecteur X.

Proposition 2.2.1. *Soit $X\ :\ (\Omega, \mathcal{F}, \mathbb{P})\ \to\ (E, \mathcal{E})$ une application et $\mathcal{C} \subseteq \mathcal{P}(E)$. Alors*

$$X^{-1}(\sigma(\mathcal{C})) = \sigma(X^{-1}(\mathcal{C}))$$

En particulier, si $\sigma(\mathcal{C}) = \mathcal{E}$, pour vérifier que X est mesurable, il suffit de vérifier que $X^{-1}(\mathcal{C}) \subseteq \mathcal{F}$.

Démonstration : On vérifie aisément que $X^{-1}(\sigma(\mathcal{C}))$ est une tribu qui contient $X^{-1}(\mathcal{C})$. Donc par minimalité

$$\sigma(X^{-1}(\mathcal{C})) \subseteq X^{-1}(\sigma(\mathcal{C}))$$

Par ailleurs, on vérifie immédiatement que

$$T = \{A \in \mathcal{E} \mid X^{-1}(A) \in \sigma(X^{-1}(\mathcal{C}))\}$$

est également une tribu, telle que $X^{-1}(T) \subseteq \sigma(X^{-1}(\mathcal{C}))$.
De plus, $\mathcal{C} \subseteq T$ car $X^{-1}(\mathcal{C}) \subseteq \sigma(X^{-1}(\mathcal{C}))$, donc $\sigma(\mathcal{C}) \subseteq T$. Ainsi,

$$X^{-1}(\sigma(\mathcal{C})) \subseteq X^{-1}(T) \subseteq \sigma(X^{-1}(\mathcal{C}))$$

D'où l'égalité. □

La somme, le produit, le quotient, le min, le max, la lim inf, la lim sup etc de variables aléatoires réelles est une variable aléatoire.

Proposition 2.2.2. *Si $(X_n) : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ est une suite de variables aléatoires, où (E, d) est un espace métrique et $\mathcal{E}$ une tribu borélienne. On suppose que, $\forall \omega \in \Omega$,*

$$X_n(\omega) \to_{n\infty} X(\omega) \in E$$

Alors, $X : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ est une variable aléatoire.

Démonstration : Il suffit de vérifier que pour tout $O \subseteq E$ ouvert, $X^{-1}(O) \in \mathcal{F}$.
On pose pour $r \in \mathbb{N}^*$

$$O_r = \{x \in O \mid d(x, E\backslash 0) > \frac{1}{r}\}$$

O_r est ouvert et $O = \bigcup_{r \geq 1} O_r$.
Alors,

$$X^{-1}(O) = \{\omega \in \Omega \mid X(\omega) \in O\} = \{\omega \in \Omega \mid \lim_{n\infty} X_n(\omega) \in O = \bigcup_{r \geq 1} O_r\}$$

D'où,

$$X^{-1}(O) = \{\omega \in \Omega \mid \exists r \geq 1,\ \exists m,\ \forall n \geq m,\ X_n(\omega) \in O_r\} = \bigcup_{r,m \geq 1} \bigcap_{n \geq m} X_n^{-1}(O_r) \in \mathcal{F}$$

Car $X_n^{-1}(O_r) \in \mathcal{F}$. □

Définition 2.2.2 (Variable aléatoire réelle étagée)**.** On dit qu'une variable aléatoire réelle X est étagée si elle ne prend qu'un nombre fini de valeurs :

$$X = \sum_{i=1}^{M} x_i \mathbb{1}_{A_i},\ A_i \in \mathcal{F}$$

Proposition 2.2.3 (Lien variable aléatoire réelle et étagée)**.** *Toute variable aléatoire réelle est limite simple de variable aléatoire étagée. De plus, si $X \geq 0$ la limité peut être choisie croissante.*

Démonstration : Quitte à prendre les parties positive et négative

$$X = X^+ - X^-,\ X^+ = X\mathbb{1}_{\geq 0} \text{ et } X^- = -X\mathbb{1}_{X \leq 0}$$

$$|X| = X^+ + X^-$$

On peut supposer directement que $X \geq 0$. Auquel cas X est la limite ponctuelle de X_n :

$$X_n(\omega) = \sum_{k=1}^{n2^n} \frac{k-1}{2^n} \mathbb{1}_{A_{k,n}}(\omega) + n\mathbb{1}_{B_n}(\omega)$$

où

$$A_{k,n} = \left\{ \omega \mid \frac{k-1}{2^n} \leq X(\omega) \leq \frac{k}{2^n} \right\}$$

$$B_n = \left\{ \omega \mid X(\omega) > n \right\}$$

On a alors

$$X_n(\omega) \to X(\omega)$$

La limite est croissante. □

2.2.2 Définitions et premières propriétés - Loi d'une variable aléatoire

Définition 2.2.3 (Loi d'une variable aléatoire)**.** Soit $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, \mathcal{E})$ une variable aléatoire. On appelle loi de X, la mesure $\mathbb{P}_X$ sur $(E, \mathcal{E})$, mesure image de $\mathbb{P}$ par X. C'est-à-dire

$$\forall A \in \mathcal{E},\ \mathbb{P}_X(A) = \mathbb{P}(X^{-1}(A)) = \mathbb{P}(X \in A) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in A\})$$

On a donc enfait que $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$.
$\mathbb{P}_X$ est une mesure de probabilité sur $(E, \mathcal{E})$.

- $\mathbb{P}_X(\emptyset) = \mathbb{P}(X^{-1}(\emptyset)) = \mathbb{P}(\emptyset) = 0$;
- $\mathbb{P}_X(E) = \mathbb{P}(X^{-1}(E)) = \mathbb{P}(\Omega) = 1$;
- $\mathbb{P}_X\left(\bigsqcup_{i=1}^{\infty} A_i\right) = \mathbb{P}\left(X \in \bigsqcup_{i=1}^{\infty} A_i\right) = \mathbb{P}\left(\bigsqcup_{i=1}^{\infty} (X \in A_i)\right) = \sum_{i=1}^{\infty} \mathbb{P}(X \in A_i) = \sum_{i=1}^{\infty} \mathbb{P}_X(A_i)$

Si $X = (X_1, ..., X_n)$ est un vecteur aléatoire à valeurs dans $= \left(\prod_{i=1}^{n} E_i, \bigotimes_{i=1}^{n} \mathcal{E}_i\right)$. Alors la loi $\mathbb{P}_X$ est appelée loi jointe des (X_i). Les lois $\mathbb{P}_{X_i}$ sont appelées lois marginales.
Si $A \in \mathcal{E}_i \subseteq \mathcal{P}(E_i)$

$$\mathbb{P}_{X_i}(A) = \mathbb{P}_X(E_1 \times E_2 \times ... \times A \times E_{i+1} \times ... \times E_n)$$

Définition 2.2.4 (Variables aléatoires de même loi)**.** On dit que deux variables aléatoires $X, Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, \mathcal{E})$ ont la même loi si $\mathbb{P}_X = \mathbb{P}_Y$. Auquel cas, on note $\mathcal{L}(X) = \mathcal{L}(Y)$ ou $X =^{\mathcal{L}} Y$.

Si $\mathcal{E} = \sigma(\mathcal{C})$ avec $\mathcal{C}$ stable par intersection. D'après le lemme des classes monotones, $X, Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, \mathcal{E})$ ont la même loi si et seulement si

$$\mathbb{P}(X \in A) = \mathbb{P}(Y \in A),\ \forall A \in \mathcal{C}$$

Définition 2.2.5 (Atomes)**.** Soit $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, \mathcal{E})$ une variable aléatoire. On appelle atome de X ou de la loi $\mathbb{P}_X$, tout $x \in E$ tel que

$$\mathbb{P}_X(\{x\}) = \mathbb{P}(X = x) > 0$$

La notion d'atome d'une variable aléatoire est un concept important.
L'application immédiate est que si on connait tous les atomes d'une variable aléatoire discrète X alors la probabilité que X prenne une valeur dans un ensemble A peut être calculée comme la somme des probabilités des atomes dans A.
On verra aussi plus tard que les atomes d'une variable aléatoire sont liés aux discontinuités de sa fonction de répartion F_X. En effet, une discontinuité à $x = a$ indique que a est un atome de X.
De manière plus abstraite, les atomes permettent de comprendre la structure de la distribution de probabilité d'une variable aléatoire. Par exemple, une variable aléatoire discrète a un ensemble fini ou dénombrable d'atomes, tandis qu'une variable aléatoire continue n'a généralement pas d'atomes.

De manière plus appliqué, dans des domaines comme la finance, l'ingénierie et la biologie, comprendre les atomes d'une variable aléatoire peut aider à modéliser et à prédire des phénomènes aléatoires. Par exemple en finance, les atomes peuvent représenter des prix de marché spécifiques qui ont une probabilité non-nulle d'être atteints.

Définition 2.2.6 (Support topologique)**.** Soit $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, \mathcal{E})$ une variable aléatoire à valeurs dans E un espace topologique. On appelle support de X, ou de sa loi, le plus petit fermé $F \subseteq E$ tel que $\mathbb{P}_X(F) = 1$.

Le support topologique permet en quelque sorte de comprendre quelles sont les valeurs "possibles" pour la variable aléatoire X et aide donc à interpréter sa distribution.
On verra que pour une variable aléatoire continue avec une fonction de densité f_X, le support topologique est souvent l'ensemble des points où $f_X(x) > 0$. Cela permet de déterminer où la densité de probabilité est concentrée. On verra aussi que le support topologique est lié à la fonction de répartition F_X. Par exemple, si F_X est constante sur un intervalle, cet intervalle n'est pas dans le support topologique de X (On ne comprend pas encore pourquoi car on a pas la définition de la fonction de répartion. Indice : $\mathbb{P}(X \in [a, b])) = F_X(b) - F_X(a)$).

De même, en finance, le support topologique peut représenter l'ensemble des prix de marché possibles pour un actif.

Exemple 2.2. On jette deux dés et on regarde la somme. $\Omega = \{1, ..., 6\}^2$, $\mathcal{F} = \mathcal{P}(\Omega)$, $\mathbb{P}$ uniforme.

$$\forall A \in \mathcal{P}(\Omega),\ \mathbb{P}(A) = \frac{\text{Card}(A)}{\text{Card}(\Omega)}$$

On considère la variable aléatoire suivante

$$\begin{array}{rccc} X & : & (\Omega, \mathcal{F}, \mathbb{P}) & \to & (\{2, ..., 12\}, \mathcal{P}(\{2, ..., 12\})) \\ & & \omega = (\omega_1, \omega_2) & \mapsto & \omega_1 + \omega_2 \end{array}$$

$$\mathbb{P}_X(\{2\}) = \mathbb{P}(X = 2) = \mathbb{P}(\omega = (\omega_1, \omega_2) \in \Omega,\ \omega_1 + \omega_2 = 2) = \mathbb{P}((\omega_1, \omega_2) = (1, 1)) = \frac{1}{36}$$

$$\mathbb{P}_X(\{3\}) = \mathbb{P}(\omega \in \{(1,2)\} \sqcup \{(2,1)\}) = \frac{2}{36}$$

On a alors,

$$\mathbb{P}_X = \frac{1}{36}(\delta_2 + \delta_{12}) + \frac{2}{36}(\delta_3 + \delta_{11}) + \frac{3}{36}(\delta_4 + \delta_{10}) + \frac{4}{36}(\delta_5 + \delta_9) + \frac{5}{36}(\delta_6 + \delta_8) + \frac{6}{36}\delta_7$$

Et on vérifie bien que la somme des fractions fait 1.

Exemple 2.3. On prend $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{B}([0,1]), \lambda)$. Soit X la variable aléatoire

$$\begin{array}{rccc} X & : & [0,1] & \to & \{0,1\}^n \\ & & x = \displaystyle\sum_{i=1}^{\infty} \frac{x_i}{2^i} & \mapsto & \{x_1, ..., x_n\} \end{array}$$

C'est-à-dire que X renvoie les n premières décimales en base 2.
Soit $\epsilon = (\epsilon_1, ..., \epsilon_n) \in \{0,1\}^n$ un n-uplet. Alors,

$$\mathbb{P}_X(\{\epsilon\}) = \mathbb{P}(\forall i \in \{1, ..., n\},\ x_i = \epsilon_i) = \mathbb{P}\left(x \in [\sum_{i=1}^{n} \frac{\epsilon_i}{2^i}, \sum_{i=1}^{n} \frac{\epsilon_i}{2^i} + \frac{1}{2^n}[\right) = \frac{1}{2^n}$$

car λ est la mesure de Lebesgue. Ainsi, $\mathbb{P}_X$ est la loi uniforme sur $\{0,1\}^n$.

2.2.3 Variables aléatoires indépendantes

Définition 2.2.7 (Indépendance)**.** Sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$, on dit que deux évènements $A, B \in \mathcal{F}$ sont indépendants si

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B)$$

On écrira $A \perp B$ lorsque A et B sont indépendants.

Plus généralement, si $(A_i)_{i \in I}$ est une collection quelconque d'évènements. On distingue deux types d'indépendance :

- Indépendance mutuelle : $\forall J \subseteq I$ fini, $\mathbb{P}\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} \mathbb{P}(A_j)$
- Indépendance deux à deux : $\forall i \neq j,\ \mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i) \times \mathbb{P}(A_j)$

Définition 2.2.8 (Familles indépendantes)**.** On dit que deux familles $\mathcal{G}, \mathcal{H} \subseteq \mathcal{F}$ sont indépendantes si

$$\forall A \in \mathcal{G}, \forall B \in \mathcal{H},\ \mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B)$$

De même et plus généralement, on peut étendre cette notion d'indépendance à une collection de familles $(\mathcal{G}_i)_{i \in I}$ en distinguant deux types d'indépendance :

- Indépendance mutuelle : $\forall J \subseteq I$ fini, $\forall A_j \in \mathcal{G}_j$, $\mathbb{P}\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} \mathbb{P}(A_j)$

- Indépendance deux à deux : $\forall i \neq j$, $\mathcal{G}_i$ est indépenante de $\mathcal{G}_j$.

Proposition 2.2.4. *Si $(\mathcal{A}_i)_{i\in I}$ est un collection de π-système, alors les $(\mathcal{A}_i)$ sont indépendants (mutuellements ou 2 à 2) si et seulement si, les $(\sigma(\mathcal{A}_\backslash))_{i\in I}$ sont indépendantes.*

Démonstration : □

Dans la suite, si X est une variable aléatoire $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, \mathcal{E})$, on note

$$\sigma(X) = X^{-1}(\mathcal{E})$$

C'est la plus petite tribu sur Ω rendant X mesurable.

Définition 2.2.9. Deux variables aléatoires X et Y telles que $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (E_1, \mathcal{E}_1)$, $Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (E_2, \mathcal{E}_2)$ sont indépendantes si les tribus $\sigma(X)$ et $\sigma(Y)$ sont indépendantes, ie

$$\forall A \in \mathcal{E}_1,\ \forall B \in \mathcal{E}_2,\ \mathbb{P}(X \in A \text{ et } Y \in B) = \mathbb{P}(X \in A) \times \mathbb{P}(Y \in B)$$

Plus généralements, on dira que les variables aléatoires d'une collection (X_i) sont mutuellements inépendantes ou 2 à 2 si les tribus associées $(\sigma(X_i))$ le sont.

Exemple 2.4. On jette deux dés et on note X_1 et X_2 les résultats :

$$(\Omega, \mathcal{F}, \mathbb{P}) = (\{1, ..., 6\}^2, \mathcal{P}(\{1, ..., 6\}^2), \text{mesure uniforme})$$

$$\Omega = \{\omega = (\omega_1, \omega_2) \mid \text{ avec } \omega_i \in \{1, ..., 6\}^2\} \text{ et } X_i(\omega) = \omega_i$$

Si $\mathbb{P}$ uniforme,

$$\forall i, j \in \{1, ..., 6\},\ \mathbb{P}(X_1 = i \text{ et } X_2 = j) = \frac{1}{36}$$

Ainsi,

$$\mathbb{P}(X_1 = i) = \sum_{j=1}^{6} \mathbb{P}(X_1 = i \text{ et } X_2 = j) = \frac{1}{6}$$

$$\mathbb{P}(X_2 = j) = \sum_{i=1}^{6} \mathbb{P}(X_1 = i \text{ et } X_2 = j) = \frac{1}{6}$$

Si bien que,

$$\mathbb{P}(X_1 = i \text{ et } X_2 = j) = \mathbb{P}(X_1 = i) \times \mathbb{P}(X_2 = j)$$

Proposition 2.2.5 (Indépendance et vecteur aléatoire)**.** *Si $X = (X_1, ..., X_n)$ est un vecteur aléatoire à valeurs dans $(\prod_{i=1}^{n} E_i, \bigotimes \mathcal{E}_i)$ alors les variables (X_i) sont mutuellement indépendantes si et seulement si, $\forall A_i \subseteq \mathcal{E}_i$ on a*

$$\mathbb{P}_X(A_1 \times ... \times A_n) = \mathbb{P}_{X_1}(A_1) \times ... \times \mathbb{P}_{X_n}(A_n)$$

C'est-à-dire,

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

Démonstration : C'est presque immédiat par définition,

$$\mathbb{P}_X(A_1\times...\times A_n) = \mathbb{P}(X \in (A_1,...,A_n)) = \mathbb{P}(X_1 \in A_1,...,X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i) = \prod_{i=1}^{n} \mathbb{P}_{X_i}(A_i)$$

□

2.2.4 Fonction de répartition

Définition 2.2.10 (Fonction de répartition)**.** On appelle fonction de répartition d'une variable aléatoire réelle la fonction $F_X : \mathbb{R} \to [0,1]$

$$F_X(x) = \mathbb{P}_X(\,]-\infty, x]) = \mathbb{P}(X \leq x),\ x \in \mathbb{R}$$

Proposition 2.2.6. *La fonction de répartition caractérise la loi d'une variable aléatoire, au sens si $F_X = F_Y$ alors $\mathbb{P}_X = \mathbb{P}_Y$.*

Démonstration : La famille $\mathcal{A} = \{]-\infty, x] \mid x \in \mathbb{R}\}$ est stable par intersection qui engendre $\mathcal{B}(\mathbb{R})$ et on conclut par le lemme des classes monotones. □

Proposition 2.2.7. *Soit F_X la fonction répartition d'une variable aléatoire réelle X. Alors*

1. *F_X est croissante ;*
2. *$\lim\limits_{-\infty} F_X(x) = 0$ et $\lim\limits_{+\infty} F_X(x) = 1$;*
3. *F_X est continue à droite limite à gauche,*

$$F_X(x^-) = \lim_{y\to x,\ y<x} F_X(y) = \mathbb{P}(X < x)$$

4. *$\mathbb{P}(X \in [a,b]) = F_X(b) - F_X(a^-)$;*
5. *Si x_0 est un atome de X alors F_X est continue en x_0.*

Démonstration : (a) Prenons $x \leq y$, alors $]-\infty, x] \subseteq]-\infty, y]$, de sorte que comme $\mathbb{P}_X$ est une mesure de probabilités

$$F_X(x) = \mathbb{P}_X(]-\infty, x]) \leq \mathbb{P}_X(]-\infty, y]) = F_X(y)$$

(b) On a

$$\lim_{-\infty} F_X(x) = \lim_{-\infty} \mathbb{P}_X(]-\infty, x]) = \lim_{\infty} \mathbb{P}_X(]-\infty, -n]) = \mathbb{P}_X\left(\bigcap_n\,]-\infty, -n]\right)$$

D'où,

$$\lim_{-\infty} F_X(x) = \mathbb{P}_X(\emptyset) = 0$$

De même,

$$\lim_{\infty} F_X(x) = \lim_{\infty} \mathbb{P}_X(]-\infty, n]) = \mathbb{P}_X\left(\bigcup_n\,]-\infty, n]\right) = \mathbb{P}_X(\mathbb{R}) = 1$$

(c) On pose $A_n =]-\infty, x+\frac{1}{n}]$ d'où $\bigcap_n A_n =]-\infty, x]$.

$$\lim_{y\to x,\ y>x} F_X(y) = \lim_{\infty} \mathbb{P}_X(A_n) = \mathbb{P}_X(\bigcap A_n) = F_X(x)$$

On pose $B_n =]-\infty, x-\frac{1}{n}]$, $\bigcup_n B_n =]-\infty, x[$

$$\lim_{y\to x,\ y<x} F_X(y) = \lim_{\infty} \mathbb{P}_X(B_n) = \mathbb{P}_X(\bigcup B_n) = \mathbb{P}_X(]-\infty, x[) = F_X(x^-)$$

(d) On a

$$\mathbb{P}(X \leq x) = \mathbb{P}(X < x) + \mathbb{P}(X = x)$$

C'est-à-dire

$$F_X(x) = F_X(x^-) + \mathbb{P}(X = x)$$

i.e. les points de discontinuités de F_X sont exactement les atomes.

□

Corollaire 2.2.1. *Une variable aléatoire réelle possède un nombre au plus dénombrable d'atomes.*

Démonstration : Une fonction monotone possède un nombre au plus dénombrable de discontinuité.
□

Proposition 2.2.8. *Soit F une fonction continue à droite limite à gauche croissante avec $\lim_{-\infty} F(x) = 0$, $\lim_{\infty} F(x) = 1$ alors F est la fonction de répartition d'une variable aléatoire réelle X.*

Démonstration : On construit X sur $(\Omega, \mathcal{F}, \mathbb{P}) = (]0,1[, \mathcal{B}(]0,1[), \text{Lebesgue } \lambda)$.
Pour $\omega \in]0,1[$, on pose

$$X(\omega) = F^{-1}(\omega) = \infty\{t \in \mathbb{R} \mid F(t) \geq \omega\}$$

X est bien définie car $\lim_{\infty} F(t) = 1$, $\lim -\infty F(t) = 0$ i.e. $\{t \in \mathbb{R} \mid F(t) \geq \omega\}$ est non-vide et minoré.
X est mesurable car F est continue à droite limite à gauche.
On a l'équivalence

$$X(\omega) \leq x \Longleftrightarrow \omega \leq F(x)$$

On a alors

$$F_X(x) = \mathbb{P}_X(]-\infty, x]) = \mathbb{P}(X \leq x) = \mathbb{P}(\omega \mid X(\omega) \leq x) = \mathbb{P}(\omega \mid \omega \leq F(x)) = F(x)$$

□

En résumé, la fonction de répartition fournit une description complète de la distribution d'une variable aléatoire. Connaitre $F_X(x)$ permet, en somme, de connaitre la probabilité que X prenne une valeur dans n'importe quel intervalle :

$$\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a)$$

On verra plus tard qu'elle est aussi utilisée pour étudier la convergence en distribution des suites de variables aléatoires. Une suite de variables aléatoires $\{X_n\}$ converge en distribution vers une variable aléatoire X si les fonctions de répartitions F_{X_n} convergent vers F_X en tout point de continuité de F_X.

En statistique, la fonction de répartition est aussi utilisée pour estimer les paramètres des distributions, tester des hypothèses et effectuer des analyses de régression. La fonction de répartition aussi d'analyser les propriétés des distributions telles que la médiane, les quartiles et d'autres quantiles. Par exemple, la médiane m est définie par $F_X(m) = 0,5$.

La fonction de répartition est aussi utilisée pour générer des échantillons de variables aléatoires dans les simulations de Monte Carlo. Par exemple, pour générer une variable aléatoire X avec une fonction de répartition F_X on peut utiliser, comme on l'a vu dans la démonstration précédente, la méthode de la transformée inverse : $X = F_X^{-1}(U)$ où U est une variable aléatoire uniforme sur $[0, 1]$.

On verra aussi plus tard qu'elle est utile dans l'étude des processus stochastiques. La fonction de répartition est utilisée pour analyser les propriétés des processus tels que les temps d'attente et de retour.

2.2.5 Notion de densité

Définition 2.2.11 (Absolue continuité)**.** Soient μ et ν deux mesures σ-finies sur $(\Omega, \mathcal{F})$. On dit que μ est absolument continue par rapport à ν et on note $\mu << \nu$ si

$$\forall A \in \mathcal{F},\ \nu(A) = 0 \Longrightarrow \mu(A) = 0$$

Intuitivement, l'absolue continuité de μ par rapport à ν signifie que μ ne "voit" pas de masse en dehors des ensembles où ν voit de la masse. En d'autres termes, si un ensemble est négligeable (de mesure nulle) pour ν, il est également négligeable pour μ.

Théorème 2.2.1 (de Radon-Nikodym)**.** *Si $\mu << \nu$ alors il existe une fonction mesurable f telle que*

$$\forall A \in \mathcal{F},\ \mu(A) = \int_A f d\nu = \int \mathbb{1}_A f d\nu$$

La fonction f est appelée densité (ou dérivée de Radon-Nikodym) de μ par rapport à ν.
Si μ et ν sont positives et μ est finie alors f est positive et $f \in L^1(\nu)$.

Démonstration : Admise (pour l'instant). □

Définition 2.2.12 (Variable aléatoire à densité)**.** On dit qu'une variable aléatoire réelle X est à densité si $\mathbb{P}_X$ est absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}$. Auquel cas la fonction $f_X \geq 0$, $f_X \in L^1$ est appelée densité de X (de la loi de X).

Autrement dit,

$$\forall A \in \mathcal{B}(\mathbb{R}),\ \mathbb{P}_X(A) = \mathbb{P}(X \in A) = \int_A f_X(x) dx$$

Et en particulier,

$$\mathbb{P}_x([a,b]) = \int_a^b f_X(x)dx$$

Exemple 2.5. On prend $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{B}([0,1]), \text{Lebesgue})$

$$\begin{array}{rccc} Y & : & (\Omega,\mathcal{F}) & \to & ([-1,0],\mathcal{B}([-1,0])) \\ & & \omega & \mapsto & \omega^2-1 \end{array}$$

$$\mathbb{P}_Y([a,b]) = \mathbb{P}(Y \in [a,b]), \text{ si } [a,b] \subseteq [-1,0]$$

Donc,

$$\mathbb{P}_Y([a,b]) = \mathbb{P}(\{\omega \,|\, \omega^2 - 1 \in [a,b]\}) = \mathbb{P}(\{\omega \,|\, \omega \in [\sqrt{a+1}, \sqrt{b+1}]\}) = \sqrt{b+1} - \sqrt{a+1} = \int_a^b \frac{1}{2\sqrt{x+1}} dx$$

Réciproquement, si f est une fonction mesurable positive d'intégrale 1 par rapport à la mesure de Lebesgue, la formule suivante définit une mesure de probabilité

$$\mathbb{P}(A) = \int_A f d\lambda$$

Proposition 2.2.9. *Si X est une variable aléatoire réelle de densité f_X. Alors,*

1. $\forall x \in \mathbb{R},\ F_X(x) = \int_{-\infty}^{x} f_X(t)dt$;
2. *F_X est continue sur $\mathbb{R}$;*
3. *Si f_X est continue en x_0 alors F_X est dérivable en x_0 et $F_X^{'}(x_0) = f(x_0)$;*
4. *Si X a pour fonction de répartition*

$$F_X(x) = \int_{-\infty}^{x} f(t)dt, \text{ avec } f \text{ mesurable positive}$$

Alors X a pour densité f.

<u>*Démonstration :*</u> (a) Par définition $F_X(x) = \mathbb{P}(X \in]-\infty, x]) = \int_{-\infty}^{x} f_X(t)dt.$

(b) La mesure de Lebesgue ne charge pas les singletons, donc $\mathbb{P}_X$ non plus. De plus, les points de discontinuité de F_X sont exactement les atomes. Donc, F_X est bien continue sur $\mathbb{R}$.

(c) Soit $\epsilon > 0$ et $\delta >> 1$ tel que $\forall |h| < \delta$,

$$|f(x_0+h) - f(x_0)| < \epsilon$$

Alors,

$$|F_X(x_0+h) - F_X(x_0) - hf_X(x_0)| = |\int_{x_0}^{x_0+h} \{f_X(t) - f_X(x_0)\}dt| \leq \epsilon|h|$$

(d) Si μ est une mesure de densité f, alors μ et $\mathbb{P}_X$ coïncident sur les ensembles $(]-\infty, x],\ x \in \mathbb{R})$ et donc sur $\mathcal{B}(\mathbb{R})$. $\square$

Si X est une variable aléatoire réelle avec F_X continue, alors X n'est pas nécessairement à densité. L'exemple classique d'une telle variable est une variable aléatoire suivant une distribution de Cantor, construite à partir de l'ensemble de Cantor. La fonction de répartition de la distribution de Cantor est appelée la fonction du diable de Cantor.
Cela vient du fait qu'une fonction continue, dérivable de dérivée L^1 n'est pas forcément l'intégrale de sa dérivée.
Cependant, si F_X est absolument continue, ie $\exists f$ intégrable telle que $F(b)-F(a) = \int_a^b f(t)dt$, alors X est à densité.

Définition 2.2.13 (Densité pour les vecteurs aléatoires)**.** Une fonction $f : \mathbb{R}^d \to \mathbb{R}$ mesurable est appelée densité si $f \geq 0$ et

$$\int_{\mathbb{R}^d} f(x)\lambda^d(dx) = 1$$

Une vecteur aléatoire $X = (X_1, ..., X_d)$ a pour loi la loi de densité f si $\forall [a_i, b_i],\ a_i < b_i,\ i = 1, .., d$

$$\mathbb{P}(X \in \prod_{i=1}^{d} [a_i, b_i]) = \int_{\prod_{i=1}^d [a_i,b_i]} f(x)\lambda^d(dx) = \int_{\prod_{i=1}^d [a_i,b_i]} f(x_1, ..., x_d)dx_1...dx_n$$

Proposition 2.2.10. *Soit X un vcteur aléatoire à valeurs dans $\mathbb{R}^d$, de densité f_X et soit φ un difféomorphisme de $\mathbb{R}^d$. Alors, $Y = \varphi(X)$ est une variable aléatoire à densité f_Y :*

$$f_Y(u) = f_X(\varphi^{-1}(u))J\varphi^{-1}(u)$$

où,

$$J\varphi^{-1}(u) = |\det(\partial_i \varphi_j^{-1}(u))|$$

Démonstration : C'est juste le changement de variable,

$$\mathbb{P}(Y \in B) = \mathbb{P}(\varphi(X) \in B) = \mathbb{P}(X \in \varphi^{-1}(B)) = \int_{\varphi^{-1}(B)} f_X(x)dx$$

En faisant le changement de variable $x = \varphi^{-1}(u)$:

$$\mathbb{P}(Y \in B) = \int_{\varphi^{-1}(B)} f_X(x)dx = \int_B f_X(\varphi^{-1}(u))J\varphi^{-1}(u)du$$

D'où la formule de la densité :

$$f_Y(u) = f_X(\varphi^{-1}(u))J\varphi^{-1}(u)$$

$\square$

Proposition 2.2.11 (Densité marginales). *Soit (X,Y) un couple de variables aléatoires réelles de densité $f_{(X,Y)}$ sur $\mathbb{R}^2$. Alors X et Y sont à densité, données par*

$$f_X(x) = \int_{\mathbb{R}} f_{(X,Y)}(x,y)dy$$

$$f_Y(y) = \int_{\mathbb{R}} f_{(X,Y)}(x,y)dx$$

Démonstration : Soit $A \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A) = \mathbb{P}(X \in A,\, Y \in \mathbb{R}) = \int_{A\times\mathbb{R}} f_{\{(X,Y)\}}(x,y)dxdy = \int_A \big(\int_{\mathbb{R}} f_{(X,Y)}(x,y)dy\big)dx$$

Où on utilise Fubini dans la dernière égalité.
On fait la même chose avec la variable aléatoire Y. □

Proposition 2.2.12 (Densité et indépendance). *Soit (X,Y) un couple de variable aléatoire de densité $f_{(X,Y)}$. Alors, X et Y sont indépendantes si et seulement si*

$$f_{(X,Y)}(x,y) = f_X(x) \times f_Y(y)$$

Démonstration : Soit $A, B \in \mathcal{B}(\mathbb{R})$ alors

$$\mathbb{P}_{(X,Y)}(A\times B) = \mathbb{P}(X \in A \text{ et } Y \in B) = \int_{A\times B} f_{(X,Y)}(x,y)dxdy$$

$$\mathbb{P}_X(A)\mathbb{P}_Y(B) = \int_A f_X(x)dx \int_B f_Y(y)dy = \int_{A\times B} f_X(x)f_Y(y)dxdy$$

C'est-à-dire,

$$X \perp Y \iff \int_{A\times B} f_{(X,Y)}(x,y)dxdy = \int_{A\times B} f_X(x)f_Y(y)dxdy,\ \forall A,B$$

□

Plus généralement, si $X = (X_1, ..., X_d)$ est à densité alors les X_i sont mutuellement indépendants

$$f_{(X_1,...,X_d)} = \prod_{i=1}^{d} f_{X_i}$$

2.3 Variables aléatoires usuelles

2.3.1 Variables aléatoires discrètes

- On dit qu'une variable aléatoire X est une variable aléatoire de Bernoulli de paramètre $p \in [0,1]$ si $X(\Omega) = \{0,1\}$ ou plus généralement un ensemble à 2 points et

$$\mathbb{P}(X=0) = 1-p,\ \mathbb{P}(X=1) = p$$

On notera $X \sim \mathcal{B}(p)$.

- On dit que X suit la loi Binomiale de paramètre $n \in \mathbb{N}^*$ et $p \in [0,1]$, $X \sim \mathcal{B}(n,p)$ si $X(\Omega) = \{0, ..., n\}$

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- On dit X suit une loi géométrique de paramètre $p \in [0,1]$, $X \sim \mathcal{G}(p)$ si $X(\Omega) = \mathbb{N}^*$ et

$$\mathbb{P}(X = k) = (1-p)^{k-1} p$$

- On dit que X suit une loi de Poisson de paramètre $\lambda > 0$, $X \sim \mathcal{P}(\lambda)$, si $X(\Omega) = \mathbb{N}$ et

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

- On dit que X suit une loi uniforme sur $\{x_1, ..., x_n\}$ si $\mathbb{P}(X = x_i) = \frac{1}{n}$, $\forall i = 1, ..., n$

$$\mathbb{P}(X \in A) = \frac{\text{Card} A}{\text{Card} \Omega}$$

2.3.2 Variables aléatoires à densité

- On dit que X est uniforme dans $[a,b]$ et on note $X \sim \mathcal{U}_{[a,b]}$ si $\mathbb{P}_X$ est à densité

$$f_X(x) = \frac{1}{b-a} \mathbb{1}_{[a,b]}(x)$$

- On dit que X suit une loi exponentielle de paramètre $\lambda > 0$, $X(\Omega) = \mathbb{R}^+$ et X a une densité

$$f_X(x) = \lambda e^{-\lambda x} \mathbb{1}_{x \geq 0}, \ \mathbb{P}(X \geq t) = e^{-\lambda t}$$

- On dit que X suit la loi normale ou gaussienne de paramètre $m \in \mathbb{R}$ et $\sigma \in [0, \infty[$ et on note $X \sim \mathcal{N}(m, \sigma^2)$ si $X(\Omega) \in \mathbb{R}$ et X a pour densité

$$f_X(x) = \frac{e^{-\frac{(x-m)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

- On dit que X suit une loi $\Gamma(n, \lambda)$ si $X(\Omega) \in \mathbb{R}^+$ et

$$f_X(x) = \frac{\lambda^n}{\Gamma(n)} e^{-\lambda x} x^{n-1} \mathbb{1}_{x \geq 0}$$

- On dit que X suit la loi de Cauchy de paramètre $\lambda > 0$ et on note $X \sim \mathcal{C}(\lambda)$ si $X(\Omega) = \mathbb{R}$ et

$$f_X(x) = \frac{1}{\pi} \frac{\lambda}{\lambda^2 + x^2}$$

2.4 Espérance et moments

On commence par rappeler la formule de transfert.
Soit $\varphi : (X, \mathcal{F}) \to (Y, \mathcal{G})$ mesurable. Si μ est une mesure sur $(X, \mathcal{F})$ alors la mesure image par φ est $\nu(B) = \mu(\varphi^{-1}(B)),\ \forall B \in \mathcal{G}$. Si $h : (Y, \mathcal{G}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mesurable alors h est ν-intégrable si et seulement si $h \circ \varphi$ est μ-intégrable

$$\int_X h \circ \varphi(x)_m u(dx) = \int_Y h(y)\nu(dy)$$

2.4.1 Espérance : définition et premières propriétés

Définition 2.4.1 (Variable aléatoire intégrable)**.** Si $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ est une variable aléatoire, on dit que X est intégrable (ou $\mathbb{P}$-intégrable) si

$$\int_\Omega |X(\omega)|\mathbb{P}(d\omega) = \int_{\mathbb{R}} |x|\mathbb{P}_X(dx) < \infty$$

Définition 2.4.2 (Espérance)**.** On définit l'éspérance d'une variable aléatoire réelle X, positive ou intégrable, que l'on note $\mathbb{E}(X)$, comm

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

Plus généralement, h est mesurable avec $h(X) \geq 0$ ou $h(X)$ $\mathbb{P}$-intégrable, on a

$$\mathbb{E}(h(X)) = \int_\Omega h(X(\omega))\mathbb{P}(d\omega) = \int_{\mathbb{R}} h(x)\mathbb{P}_X(dx)$$

Si $X = (X_1, ..., X_n)$ est un vecteur aléatoire $h : \mathbb{R}^n \to \mathbb{R}$ mesurable

$$\mathbb{E}(h(X_1, ..., X_n)) = \int_{\mathbb{R}^n} h(x)\mathbb{P}_{(X_1,...,X_n)}(dx_1...dx_n)$$

Les intégrales ci-dessus sont à considérer au sens de Lebesgue

- Si $X = \mathbb{1}_A$, $\mathbb{E}(X) = \mathbb{P}(A)$;
- Si $X = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i}$, $\mathbb{E}(X) = \sum_{i=1}^{n} a_i \mathbb{P}(A_i)$;
- Si $X \geq 0$, $\mathbb{E}(X) = \sup\{\mathbb{E}(Y) \mid Y \text{ étagée positive, } Y \leq X\}$;
- Si $\mathbb{E}(|X|) < \infty$, $\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-)$.

Concrètement,

1. Si X est discrète $X(\Omega) = \{x_1, ..., x_n, ...\}$ et $\mathbb{P}(X = x_i) = p_i \in [0, 1]$.

$$\mathbb{E}(X) = \sum_{i=1}^{...} x_i \mathbb{P}(X = x_i) = \sum_{i=1}^{...} x_i p_i$$

$$\mathbb{E}(h(X)) = \sum_{i=1}^{...} h(x_i)p_i = \sum_{i=1}^{...} h(x_i)\mathbb{P}(X = x_i)$$

2. Si X est à densité f_X, $\mathbb{P}_X(dx) = f_X(x)dx$

$$\mathbb{E}(X) = \int x f_X(x)dx$$

$$\mathbb{E}(h(X)) = \int h(x) f_X(x)dx$$

Définition 2.4.3 (Espérance vecteur aléatoire)**.** Si $X = (X_1, ..., X_n)$ est un vecteur aléatoire avec $\mathbb{E}(|X_i|) < \infty$ pour $i = 1, ..., n$ alors on définit

$$\mathbb{E}(X) = (\mathbb{E}(X_1), ..., \mathbb{E}(X_n))$$

On dit que X est centrée si $\mathbb{E}(X_i) = 0$ pour tout i.

Proposition 2.4.1. *Si X et Y sont deux variables aléatoires réelles (positives ou intégrables)*

1. *Si $X \leq Y$ p.s. alors $\mathbb{E}(X) \leq \mathbb{E}(Y)$;*
2. *$\forall \lambda, \mu$, $\mathbb{E}(\lambda X + \mu Y) = \lambda \mathbb{E}(X) + \mu \mathbb{E}(Y)$;*
3. *$|\mathbb{E}(X)| \leq \mathbb{E}(|X|)$;*
4. *Si $X \geq 0$ p.s. et $\mathbb{E}(X) = 0$ alors $X = 0$ p.s.*

Démonstration : Voir cours d'intégration. □

On remarque d'ailleurs, en vertu de ce même cours d'intégration, que l'espérance n'est autre qu'une intégrale par rapport à $\mathbb{P}$ donc les théorèmes classiques s'appliquent

- Convergence monotone : Si X_n, suite positive, croit vers X alors

$$\mathbb{E}(\liminf_{n\infty} X_n) = \lim_{n\infty} \mathbb{E}(X_n)$$

- Lemme de Fatou : Si $X_n \geq 0$

$$\mathbb{E}(\liminf_{n\infty} X_n) \leq \liminf_{n\infty} \mathbb{E}(X_n)$$

- Convergence dominée : Si $X_n(\omega) \to X(\omega)$, $\forall \omega \in \Omega$ et $|X_n(\omega)| \leq Y(\omega)$ avec Y intégrable, alors

$$\mathbb{E}(X) = \mathbb{E}(\lim_{n\infty} X_n) = \lim_{n\infty} \mathbb{E}(X_n)$$

Proposition 2.4.2 (Inégalité de Markov)**.** *Si $t > 0$, alors*

$$\mathbb{P}(|X| > t) \leq \frac{\mathbb{E}(|X|)}{t}$$

Démonstration : $\mathbb{E}(|X|) = \mathbb{E}(|X|\mathbb{1}_{|X|>t}) + \mathbb{E}(|X|\mathbb{1}_{|X|\leq t}) \geq t\mathbb{E}(\mathbb{1}_{|X|>t}) = t\mathbb{P}(|X| > t)$,
car $\mathbb{E}(|X|\mathbb{1}_{|X|\leq t}) \geq 0$. □

Proposition 2.4.3 (Inégalité de Jensen)**.** *Soit X une variable aléatoire réelle, φ une fonction convexe telle que X et $\varphi(X)$ sont intégrable*

$$\mathbb{E}(\varphi(X)) \geq \varphi(\mathbb{E}(X))$$

Démonstration : $\varphi(x) = \sup\{f(x)_|\ f \text{ affine},\ f \leq \varphi\}$ et on conclut par positivité et linéarité de l'espérance. □

Par exemple, $\mathbb{E}(X^2) \geq \mathbb{E}(X)^2$.

Proposition 2.4.4. *Soit X une variable aléatoire positive, alors*

$$\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X > t)dt = \int_0^\infty (1 - F_X(t))dt$$

Si $X(\Omega) \subseteq \mathbb{N}$ alors $\mathbb{E}(X) = \sum_{k=0}^{\infty} \mathbb{P}(X > k)$.

Démonstration : D'après Fubini-Tonelli

$$\int_0^\infty \mathbb{P}(X > t)dt = \int_0^\infty \mathbb{E}(\mathbb{1}_{X>t})dt = \mathbb{E}(\int_0^\infty \mathbb{1}_{X>t}dt)$$

D'où,

$$\int_0^\infty \mathbb{P}(X > t)dt = \mathbb{E}(\int_0^X dt) = \mathbb{E}(X)$$

□

2.4.2 Moments d'ordre supérieurs

Définition 2.4.4 (Moment d'ordre p)**.** On dit qu'une variable aléatoire réelle X admet un moment d'ordre $p > 0$ si

$$\mathbb{E}(|X|^p) = \int_\Omega |X(\omega)|^p \mathbb{P}(d\omega) = \int_{\mathbb{R}} |x|^p \mathbb{P}_X(d\omega) < \infty$$

On écrit $||X||_p = \mathbb{E}(|X|^p)^{\frac{1}{p}}$ et on note $L^p(\Omega, \mathcal{F}, \mathbb{P})$ l'espace des variables aléatoires admettant un moment d'ordre p (on identifie les variables aléatoires qui sont égales p.s., ou encore qui coïncident en dehors d'un ensemble $\mathbb{P}$-négligeable). De plus,

$$L^\infty(\Omega, \mathcal{F}, \mathbb{P}) = \{X \mid \exists c > 0,\ \mathbb{P}(|X| > c) = 0\}$$

Les inégalités vu en intégration "s'étendent" dans ce cadre probabiliste.

- Cauchy-Schwarz et Hölder : Si $p \geq 1$, $\frac{1}{p} + \frac{1}{q} = 1$ si $X \in L^p$, $Y \in L^q$

$$||XY||_1 = \mathbb{E}(|XY|) \leq \mathbb{E}(|X|^p)^{\frac{1}{p}} \mathbb{E}(|Y|^q)^{\frac{1}{q}} = ||X||_p ||Y||_q$$

  En particulier, si $p = 2$,

$$\mathbb{E}(|XY|) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}$$

- Minkowski : Si $p \geq 1$, $X, Y \in L^p$

$$||X + Y||_p \leq ||X||_p + ||Y||_p$$

Définition 2.4.5 (Variance)**.** Si $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, la variance de X est la quantité suivante

$$\mathrm{Var}(X) = \mathbb{E}(|X - \mathbb{E}(X)|^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

C'est l'écart quadratique à la moyenne.

L'écart type est $\sigma(X) = \sqrt{\mathrm{Var}(X)} = ||X - \mathbb{E}(X)||_2$. Si $X, Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, on définit la covariance comme

$$\mathrm{Cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) \in \mathbb{R}$$

C'est la version polarisée de la variance.

Proposition 2.4.5 (Propriétés de la variance)**.** *On a les propriétés suivantes :*

1. $Var(X) \geq 0$;
2. $Var(\lambda X) = \lambda^2 Var(X)$;
3. $Var(X + \alpha) = Var(X)$ *si* $\alpha \in \mathbb{R}$;
4. $Var(X) = 0 \Longrightarrow X$ *est constante p.s.,* $X = \mathbb{E}(X)$ *p.s.* ;
5. $Var(X + Y) = Var(X) + Var(Y) + 2\,Cov(X, Y)$.

Proposition 2.4.6 (Inégalité de Bienaymé-Tchebychev)**.** *Si* $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ *et* $t > 0$ *alors*

$$\mathbb{P}(|X - \mathbb{E}(X)| > t) \leq \frac{Var(X)}{t^2}$$

Démonstration : C'est l'inégalité de Markov

$$\mathbb{P}(|X - \mathbb{E}(X)| > t) = \mathbb{P}(|X - \mathbb{E}(X)|^2 > t^2) \leq \frac{\mathbb{E}(|X - \mathbb{E}(X)|^2)}{t^2} = \frac{\mathrm{Var}(X)}{t^2}$$

□

2.4.3 Moments des variables usuelles

- Si $X \sim \mathcal{U}_{\{x_1,...,x_n\}}$ alors
$$\mathbb{E}(X) = \sum_{k=1}^{n} x_k \frac{1}{n} = \frac{1}{n} \sum_{k=1}^{n} x_k$$
La moyenne arithmétique.

- Si $X \sim \mathcal{B}(p)$, i.e. $\mathbb{P}(X = 0) = 1 - p,\ \mathbb{P}(X = 1) = p$
$$\mathbb{E}(X) = 0.\mathbb{P}(X = 0) + 1.\mathbb{P}(X = 1) = p$$
$$\mathbb{E}(X^2) = 0^2.\mathbb{P}(X = 0) + 1^2.\mathbb{P}(X = 1) = p$$
D'où,
$$\mathrm{Var}(X) = p - p^2 = p(1 - p)$$

- Si $X \sim \mathcal{B}(n, p)$, alors comme une loi binomiale est une succession de n épreuve de Bernouilli, alors
$$X = X_1 + ... + X_n$$
où les X_i sont des variables aléatoires suivant une $\mathcal{B}(p)$. Par linéarité de l'espérance
$$\mathbb{E}(X) = \mathbb{E}(X_1 + ... + X_n) = \sum_{k=1}^{n} \mathbb{E}(X_k) = np$$
D'autre part, puisque les X_i sont indépendantes
$$\mathrm{Var}(X) = \mathrm{Var}(X_1 + ... + X_n) = \sum_{k=1}^{n} \mathrm{Var}(X_k) = np(1 - p)$$
Ceci nécessite l'indépendance des X_i pour que la covariance soit nulle.

- Si $X \sim \mathcal{G}(p)$,
$$\mathbb{P}(X = k) = p(1 - p)^{k-1},\ k \in \mathbb{N}^*$$
Puis, par dérivation
$$\mathbb{E}(X) = \sum_{k=1}^{\infty} kp(1 - p)^{k-1} = \frac{1}{p}$$
$$\mathrm{Var}(X) = \sum_{k=1}^{\infty} k^2 p(1 - p)^{k-1} = \frac{1 - p}{p^2}$$

- Si $X \sim \mathcal{P}(\lambda)$,
$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$
$$\mathbb{E}(X) = \sum_{k=0}^{\infty} ke^{-\lambda} \frac{\lambda^k}{k!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \lambda$$
$$\mathrm{Var}(X) = \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} = \lambda$$

- Si $X \sim \mathcal{U}_{[a,b]}$,

$$f_X(x) = \frac{1}{b-a}\mathbb{1}_{[a,b]}(x)$$

$$\mathbb{E}(X) = \int_a^b x\frac{1}{b-a}dx = \frac{1}{b-1}[\frac{x^2}{2}]_a^b = \frac{1}{2}\frac{b^2-a^2}{b-a} = \frac{a+b}{2}$$

De même,

$$\mathrm{Var}(X) = \frac{(b-a)^2}{12}$$

- Si $X \sim \mathcal{E}(\lambda)$,

$$f_X(x) = \lambda e^{-\lambda x}\mathbb{1}_{x\geq 0}$$

Par intégration par parties,

$$\mathbb{E}(X) = \int_0^\infty x\lambda e^{-\lambda x}dx = [-e^{-\lambda x}x]_0^\infty + \int_0^\infty e^{-\lambda x}dx = \frac{1}{\lambda}$$

$$\mathrm{Var}(X) = \frac{1}{\lambda^2}$$

- Si $X \sim \mathcal{N}(m, \sigma^2)$ alors on a

$$f_X(x) = \frac{e^{-\frac{(x-m)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

Si bien que

$$\mathbb{E}(X) = m \text{ et } \mathrm{Var}(X) = \sigma^2$$

- Si $X \sim \mathcal{C}(1)$,

$$f_X(x) = \frac{1}{\pi}\frac{\lambda}{\lambda^2+x^2}$$

$$\mathbb{E}(|X|) = \frac{1}{\pi}\int_{\mathbb{R}}\frac{|x|}{1+x^2}dx = +\infty$$

2.4.4 Espérance et identification de loi : Identification et fonctions tests

Théorème 2.4.1. *Soit X un vecteur aléatoire à valeurs dans $\mathbb{R}^d$ et μ une mesure de probabilité sur $\mathbb{R}^d$. Alors X suit la loi μ, $X \sim \mu$ si et seulement si pour tout h continue à support compact (ou $\mathcal{C}^\infty$ à support compact)*

$$\mathbb{E}(h(X)) = \int_{\mathbb{R}^d} h(x)\mu(dx)$$

Démonstration : Le sens direct est évident, si $\mathbb{P}_X = \mu$ alors

$$\mathbb{E}(h(X)) = \int h(x)\mathbb{P}_X(dx) = \int h(x)\mu(dx)$$

Réciproquement, si pour tout h continue à support compact, on a l'équation ci-dessus, fixons K compact de $\mathbb{R}^d$ et posons

$$h_n(x) = \frac{d(x, \overline{O_n})}{d(x,K) + d(x,\overline{O_n})},\ O_n = \{x \mid d(x,K) < \frac{1}{n}\}$$

Alors, (h_n) est une suite décroissante de fonctions continue à support compact positive

$$\mathbb{1}_K(x) \leq h_n(x) \leq \mathbb{1}_{O_n}(x)$$

Ainsi,

$$\lim_{n\infty} h_n(x) = \mathbb{1}_K(x)$$

Par convergence dominée,

$$\mathbb{P}_X(K) = \lim_{n\infty} \int h_n(x)\mathbb{P}_X(dx) = \lim_{n\infty} \int h_n(x)\mu(dx) = \mu(K)$$

Si bien que, par le lemme des classes monotones, on conclut

$$\mathbb{P}_X = \mu$$

□

Ce théorème est très utile pour reconnaitre et caractériser les lois.

Exemple 2.6. Soit $X \sim \mathcal{C}(1)$ et $Y = X^+ = X\mathbb{1}_{X\geq 0}$,

$$f_X(x) = \frac{1}{\pi(1+x^2)}$$

$$\mathbb{E}(h(Y)) = \mathbb{E}(h(X^+)) = \int h(x^+)\mathbb{P}_X(dx) = \int h(x^+)f_X(x)dx = \int h(x^+)\frac{dx}{\pi(1+x^2)}$$

D'où,

$$\mathbb{E}(h(Y)) = h(0)\mathbb{P}(X \leq 0) + \int_0^\infty h(x)\frac{dx}{\pi(1+x^2)} = \frac{h(0)}{2} + \int_0^\infty h(x)\frac{dx}{\pi(1+x^2)}$$

C'est-à-dire,

$$\mathbb{P}_Y = \frac{1}{2}\delta_0 + \frac{1}{\pi(1+x^2)}\mathbb{1}_{x>0}dx$$

2.4.5 Espérance et identification de loi : Espérance et indépendance

Proposition 2.4.7. *Soit $X = (X_1, ..., X_n)$ un vecteur aléatoire à valeurs dans $\mathbb{R}^n$. Il y a équivalence :*

1. *Les variables aléatoires (X_i) sont mutuellement indépendantes ;*
2. *Pour toutes fonctions mesurables bornées (ou positives) $h : \mathbb{R} \to \mathbb{R}$*

$$\mathbb{E}(h_1(X_1)...h_n(X_n)) = \prod_{i=1}^n \mathbb{E}(h_i(X_i))$$

Démonstration : On commence par le sens direct. On a vu que les (X_i) sont mutuellement indépendantes si et seulement si

$$\mathbb{P}_X = \mathbb{P}_{(X_1,...,X_n)} = \mathbb{P}_{X_1} \otimes ... \otimes \mathbb{P}_{X_n}$$

Si les h_i sont mesurables bornées, on a alors par Fubini

$$\mathbb{E}(h(X_1)...h(X_n)) = \int_{\mathbb{R}^n} h_1(x_1)...h_n(x_n)\mathbb{P}_X(dx_1...dx_n) = \int_{\mathbb{R}} h(x_1)\mathbb{P}_{X_1}(dx_1)... \int_{\mathbb{R}} h_n(x_n)\mathbb{P}_{X_n}(dx_n)$$

$$\mathbb{E}(h(X_1)...h(X_n)) = \mathbb{E}(h_1(X_1))...\mathbb{E}(h_n(X_n))$$

Réciproquement, si l'égalité du dessus est vraie, en prenant $h_i = \mathbb{1}_{A_i}$

$$\mathbb{P}(X_1 \in A_1, ..., X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i)$$

□

Corollaire 2.4.1. *Soient X, Y des variables aléatoires de $L^2(\Omega, \mathcal{F}, \mathbb{P})$ indépendantes alors $Cov(X, Y) = 0$ et*

$$Var(X + Y) = Var(X) + Var(Y)$$

Démonstration : On rappelle que

$$\mathrm{Cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))$$

Si $X \perp Y$ alors

$$\mathrm{Cov}(X, Y) = \mathbb{E}(X - \mathbb{E}(X))\mathbb{E}(Y - \mathbb{E}(Y)) = 0.0 = 0$$

□

C'est avec cette propriété que l'on démontre les formules de l'espérance et de la variance d'une loi binomiale à partir de celles d'une loi de Bernoulli.

Cependant, la réciproque est fausse comme le montre l'exemple suivant.

Exemple 2.7. Soit $U \sim \mathcal{U}_{[-1,1]}$ uniforme, $V = U^2$. Alors $\mathbb{E}(U) = 0$.

$$\mathbb{E}(V) = \mathbb{E}(U^2) = \frac{1}{2}\int_{-1}^{1} u^2 du = \frac{1}{3}$$

$$\mathrm{Cov}(U, V) = \mathbb{E}((U - \mathbb{E}(U)(V - \mathbb{E}(V))) = \mathbb{E}(U(V - \frac{1}{3})) = \mathbb{E}(U^3) - \frac{1}{3}\mathbb{E}(U) = 0$$

Mais U et V ne sont pas indépndants !

$$0 = \mathbb{P}(|U| < \frac{1}{2}, V > \frac{1}{2}) \neq \mathbb{P}(|U| > \frac{1}{2})\mathbb{P}(V > \frac{1}{2}) > 0$$

2.4.6 Espérance et identification de loi : Le problème des moments

Ce problème consiste à se demander sur la donnée des $\mathbb{E}(X^p)$, $p \geq 0$ permet d'identifier $\mathbb{P}_X$. En fait, il y a deux questions sous-jacentes :

- Existence : Si $(m_p)_{p\geq 0}$, existe-t-il une variable aléatoire X telle que $\mathbb{E}(X^p) = m_p$ pour tout $p \geq 0$?

- Unicité : Si X et Y sont telles que $\mathbb{E}(X^p) = \mathbb{E}(Y^p)$ pour tout $p \geq 0$, a-t-on $\mathbb{P}_X = \mathbb{P}_Y$?

La réponse à ces questions dépend du support de la loi cible.

Définition 2.4.6 (Suite complètement monotone)**.** Soit (m_p) une suite réelle, on dit que (m_p) est complètement monotone si

$$\forall k, p \geq 0,\ (-1)^k(\Delta^k m)_p \geq 0$$

où $(\Delta m)_p = m_{p+1} - m_p$.

Par exemple, $(\Delta^2 m)_p = (m_{p+2} - m_{p+1}) - (m_{p+1} - m_p) = m_{p+2} - 2m_{p+1} + m_p$.
Et, $(\Delta^4 m)_p = m_{10} - 4m_9 + 6m_8 - 4m_7 + m_6$.

Théorème 2.4.2 (Moments de Hausdorff)**.** *La suite $(m_p)_{p\geq 0}$ est la suite des moments d'une mesure μ à support dans $[0,1]$ i.e. $m_p = \int_0^1 x^p \mu(dx)$ si et seulement si la suite (m_p) est complètement monotone.*

Démonstration : Le caractère nécessaire de la complète monotonie est clair si l'on remarque

$$(-1)^k(\Delta^k m)_p = \int_0^1 x^p(1-x)^k \mu(dx)$$

On admet l'autre implication. □

Théorème 2.4.3. *La suite (m_p) est la suite des moments d'une mesure à support non bornée dans $\mathbb{R}$ si et seulement si la matrice infinie*

$$A = \begin{pmatrix} m_0 & m_1 & m_2 & \cdots \\ m_1 & m_2 & m_3 & \cdots \\ m_2 & m_3 & m_4 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix}$$

est définie positive.
C'est-à-dire, si pour toute suite $(a_j) \in (\mathbb{C}^{\mathbb{N}})^*$ nulle à partir d'un certain rang,

$$\sum_{j,k\geq 0} m_{j+k} a_j \overline{a_k} > 0$$

Démonstration : La condition ci-dessus est nécessaire car

$$\sum_{j,k\geq 0} m_{j+k} a_j \overline{a_k} = \int_{\mathbb{R}} \left| \sum_{j\geq 0} a_j x^j \right|^2 \mu(dx) \geq 0$$

On admet l'autre implication. □

Proposition 2.4.8. *Si μ est à support compact dans $[0,1]$ alors μ est caractérisée de manière unique par ses moments*

$$m_p = \int x^p \mu(dx)$$

Démonstration : On considère deux mesures μ et ν avec

$$\int_0^1 x^p \mu(dx) = \int_0^1 x^p \nu(dx),\ \forall p \geq 0$$

Alors, par linéarité, on a pour tout $Q \in \mathbb{R}[X]$,

$$\int_0^1 Q(x)\mu(dx) = \int_0^1 Q(x)\nu(dx)$$

Maintenant, soit f continue sur $[0,1]$, par le théorème de Stone-Weierstrass, pour $\epsilon > 0$ il existe $Q_\epsilon \in \mathbb{R}[X]$ tel que

$$||f(x) - Q_\epsilon(x)||_\infty \leq \epsilon$$

On a alors

$$\int f(x)\mu(dx) - \int f(x)\nu(dx) = \int (f(x) - Q_\epsilon(x) + Q_\epsilon(x))\mu(dx) - \int (f(x) - Q_\epsilon(x) + Q_\epsilon(x))\nu(dx)$$

Par linéarité on sépare les intégrales et par ce qui précède

$$\int Q_\epsilon \mu(dx) - \int Q_\epsilon(x)\nu(dx) = 0$$

Ainsi,

$$\left| \int f(x)\mu(dx) - \int f(x)\nu(dx) \right| \leq \left| \int (f(x) - Q_\epsilon(x))\mu(dx) \right| + \left| \int (f(x) - Q_\epsilon(x))\nu(dx) \right|$$

$$\leq \int ||f(x) - Q_\epsilon(x)||_\infty \mu(dx) + \int ||f(x) - Q_\epsilon(x)||_\infty \nu(dx) \leq \epsilon\mu([0,1]) + \epsilon\nu([0,1])$$

$$\leq 2\epsilon$$

C'est-à-dire

$$\int f(x)\mu(dx) = \int f(x)\nu(dx),\ \forall f \in \mathcal{C}^0([0,1])$$

D'où $\mu = \nu$. □

Lorsque μ n'est pas à support compact, il n'y a en général pas unicité.

Exemple 2.8. On veut illustrer ce contre-exemple en considérant deux mesures différentes qui ne sont pas à support compact mais ayant les mêmes moments.

On considère $X \sim \mathcal{N}(0,1)$. Alors,

$$\mathbb{E}(X^n) = 0, \text{ si } n \text{ est impair}$$

Maintenant, $\mathbb{E}(X^2) = 1$ et

$$\mathbb{E}(X^{2(n+1)}) = 2\int_0^\infty x^{2(n+1)} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx = 2\bigg(\int_0^\infty x^{2n+1} x \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx \bigg)$$

Par une intégration par parties

$$\mathbb{E}(X^{2(n+1)}) = 2\bigg(\bigg[-x^{2n+1} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \bigg]_0^\infty + \frac{2n+1}{\sqrt{2\pi}} \int_0^\infty x^{2n} e^{-\frac{x^2}{2}} dx \bigg) = (2n+1)\mathbb{E}(X^{2n})$$

Par récurrence,

$$\mathbb{E}(X^{2n}) = \frac{(2n!)}{2^n n!}$$

Maintenant, on considère la variable aléatoire $Y = e^X$. On appelle cela une variable aléatoire log-normale. On montre que

$$\mathbb{E}(Y^n) = e^{\frac{n^2}{2}}$$

Après cela, on considère la famille de variable aléatoire $(Y_a \mid |a| \leq 1)$ dont les densités sont données par

$$f_{Y_a}(x) = f_Y(x)(1 + a\sin(2\pi \ln(x))),\ x > 0$$

On calcule les moments des Y_a par un calcul d'intégrales et on trouve que

$$\mathbb{E}(Y_a^n) = \mathbb{E}(Y^n)$$

Ce qui termine la preuve de la nécessité de la compacité du support.

2.5 Transformées exponentielles

2.5.1 Fonction caractéristique

Définition 2.5.1 (Fonction caractéristique)**.** Soit $X = (X_1, ..., X_n)$ un vecteur aléatoire. On définit sa fonction caractéristique $\varphi_X : \mathbb{R}^d \to \mathbb{C}$

$$\varphi_X(t) = \mathbb{E}(e^{it.X}) = \mathbb{E}\bigg(\exp(i \sum_{k=1}^d t_k X_k) \bigg),\ \forall t = (t_1, ..., t_d)$$

On voit que φ_X est toujours bien définie car

$$|e^{itX}| \leq 1 \text{ et } e^{itX} \in L^1(\Omega, \mathcal{F}, \mathbb{P})$$

On a de plus un lien direct avec la transformée de Fourier de $\mathbb{P}_X$:

$$\varphi_X(t) = \int e^{it.X} \mathbb{P}_X(dx)$$

On voit maintenant quelques exemples de calculs de fonctions caractéristiques pour se familiariser avec cet objet.

Exemple 2.9. Si $X \sim \mathcal{B}(p)$

$$\mathbb{E}(e^{itX}) = e^{it.0}\mathbb{P}(X=0) + e^{it.1}\mathbb{P}(X=1) = (1-p) + pe^{it}$$

Exemple 2.10. Si $U_\lambda \sim \text{Laplace}(\lambda)$

$$f_{U_\lambda}(x) = \frac{\lambda}{2}e^{-\lambda|x|}$$

$$\mathbb{E}(e^{itU_\lambda}) = \int_{\mathbb{R}} e^{itx}\frac{\lambda}{2}e^{-\lambda|x|}dx = \frac{\lambda}{\lambda - it} + \frac{\lambda}{\lambda + it} = \frac{\lambda^2}{\lambda^2 + t^2}$$

En particulier, pour $\lambda = 1$

$$\varphi_{U_1}(t) = \frac{1}{1+t^2}$$

Plus généralement,

$$\varphi_{U_\lambda}(\lambda t) = \frac{1}{1+t^2}$$

Exemple 2.11. Si $X \sim \mathcal{N}(m, \sigma^2)$,

$$X - m \sim \mathcal{N}(0, \sigma^2) \text{ et } \frac{X-m}{\sigma} \sim \mathcal{N}(0,1)$$

On prend $X \sim \mathcal{N}(0,1)$.

$$\varphi_X(t) = \mathbb{E}(e^{itX}) = \int_{\mathbb{R}} e^{itx}e^{-\frac{x^2}{2}}\frac{dx}{\sqrt{2\pi}} = e^{-\frac{t^2}{2}}\int_{\mathbb{R}} e^{-\frac{1}{2}(x-it)^2}\frac{dx}{\sqrt{2\pi}} = e^{-\frac{t^2}{2}}$$

car la dernière intégrale vaut 1.
On peut aussi dériver sous le signe somme et trouver

$$\begin{cases} \varphi_X'(t) = -t\varphi_X(t) \\ \varphi_X(0) = 1 \end{cases}$$

Et on retrouve alors le même résultat.
La gaussienne $\mathcal{N}(0,1)$ est un point fixe de la transformée de Fourier.

$$X \sim \mathcal{N}(0, \sigma^2) \Longrightarrow \varphi_X(t) = e^{-\frac{\sigma^2 t^2}{2}}$$

Théorème 2.5.1. *La fonction caractéristique caractérise la loi, i.e. si $\varphi_X = \varphi_Y$ alors $\mathbb{P}_X = \mathbb{P}_Y$.*

Démonstration : Soit X une variable aléatoire de loi $\mathbb{P}_X$ et $U = U_\lambda$ une variable aléatoire indépendante de X de loi Laplace(λ)

$$f_U(x) = \frac{\lambda}{2}e^{-\lambda|x|}$$

$$\varphi_U(t) = \frac{\lambda^2}{\lambda^2 + t^2}$$

En particulier,

$$\varphi_U(\lambda t) = \frac{1}{1+t^2}$$

Maintenant soit $h : \mathbb{R} \to \mathbb{R}$ continue à support compact. Par convergence dominée, pour tout x

$$h(x) = \lim_{\lambda \to 0} \int_{\mathbb{R}} h(x - \lambda t) \frac{dt}{\pi(1+t^2)} = \lim_{\lambda \to 0} \frac{1}{\pi} \int_{\mathbb{R}} h(x - \lambda t) \mathbb{E}(e^{i\lambda t U}) dt$$

Toujours par convergence dominée

$$h(X) = \lim_{\lambda \to 0} \frac{1}{\pi} \int_{\mathbb{R}} h(X - \lambda t) \mathbb{E}(e^{i\lambda t U}) dt$$

$$\mathbb{E}(h(X)) = \mathbb{E}\left(\lim_{\lambda \to 0} \frac{1}{\pi} \int_{\mathbb{R}} h(X - \lambda t) \mathbb{E}(e^{i\lambda t U}) dt \right) = \lim_{\lambda \to 0} \frac{1}{\pi} \mathbb{E}\left(\int_{\mathbb{R}} h(X - \lambda t) \mathbb{E}(e^{i\lambda t U}) dt \right)$$

car l'intégrale à l'intérieure est bornée. Puis par Fubini (l'espérance c'est juste une intégrale) :

$$\mathbb{E}(h(X)) = \lim_{\lambda \to 0} \frac{1}{\pi} \int_{\mathbb{R}} \mathbb{E}(h(X - \lambda t)) \mathbb{E}(e^{i\lambda t U}) dt$$

Par indépendance de X et U :

$$\mathbb{E}(h(X)) = \lim_{\lambda \to 0} \frac{1}{\pi} \int_{\mathbb{R}} \mathbb{E}\big(h(X - \lambda t) e^{i\lambda t U}\big) dt$$

Puis en faisant les changements de variables successifs $s = \lambda t$ et $X - s = v$:

$$\mathbb{E}(h(X)) = \lim_{\lambda \to 0} \frac{1}{\pi} \int_{\mathbb{R}} \mathbb{E}\big(h(X - s) e^{isU}\big) \frac{dt}{\lambda} = \lim_{\lambda \to 0} \frac{1}{\pi} \int_{\mathbb{R}} h(v) \mathbb{E}\big(e^{i(X-v)U}\big) \frac{dv}{\lambda}$$

On remarque que

$$\mathbb{E}\big(e^{i(X-v)U}\big) = \mathbb{E}\big(e^{-ivU} \varphi(X)(U)\big)$$

Finalement

$$\mathbb{E}(h(X)) = \lim_{\lambda \to 0} \frac{1}{\pi} \int_{\mathbb{R}} h(v) \mathbb{E}\big(e^{-ivU} \varphi_X(U)\big) \frac{dv}{\lambda}$$

Et donc, $\mathbb{E}(h(X))$ ne dépend de X que via φ_X.
Ainsi, si $\varphi_X = \varphi_Y$, on déduit $\mathbb{E}(h(X)) = \mathbb{E}(h(Y))$.
Par caractérisation par les fonctions continues à support compact, on conclut que $\mathbb{P}_X = \mathbb{P}_Y$. □

On voit aussi que le théorème du théorème d'inversion de Fourier (vu juste après). En effet, si $\mathcal{F}$ est la transformée de Fourier et $\mathcal{F}^{-1}$ son inverse, on a de fait

$$\mathbb{P}_X = \mathcal{F}^{-1}(\varphi_X)$$

Ainsi, si $\varphi_X = \varphi_Y$ alors $\mathbb{P}_X = \mathbb{P}_Y$.

Proposition 2.5.1 (Propriétés de la fonction caractéristique)**.** *Soit X un vecteur aléatoire et φ_X sa fonction caractéristique*

1. $|\varphi_X(t)| \leq 1$;

2. $\varphi_X(-t) = \overline{\varphi_X(t)}$;

3. $\varphi_X(0) = 1$;

4. φ_X *est uniformément continue ;*

5. φ_X *est de type positif, c'est-à-dire*

$$\forall n,\ \forall t_1, ..., t_n \in \mathbb{R}^d,\ \forall z_1, ..., z_n \in \mathbb{C},\ \sum_{j,k=1}^{n} \varphi_X(t_j - t_k) z_k \overline{z_j} \geq 0$$

Démonstration : (a) Ok car $|e^{itX}| \leq 1$;

(b) Ok par propriété de l'exponentielle complexe ;

(c) Immédiat ;

(d) On forme la différence

$$\varphi_X(t+h) - \varphi_X(t) = \mathbb{E}\big(e^{i(t+h)X} - e^{itX}\big) = \mathbb{E}\big(e^{i(t+\frac{h}{2})X}(e^{i\frac{h}{2}X} \sin(\frac{hX}{2}))\big)$$

Ainsi,

$$|\varphi_X(t+h) - \varphi_X(t)| \leq 2\mathbb{E}(|\sin(\frac{hX}{2})|)$$

Le membre de droite est indépendant de t et tend vers 0 lorsque $h \to 0$, d'où l'uniforme continuité ;

(e) Immédiat en remarquant que

$$\sum_{j,k=1}^{n} \varphi_X(t_j - t_k) z_k \overline{z_j} = \mathbb{E}\left(\left|\sum_{k=1}^{n} e^{it_k X} z_k\right|^2\right) \geq 0$$

□

Théorème 2.5.2 (de Bochner-Herglotz)**.** *Soit $\varphi : \mathbb{R}^d \to \mathbb{C}$ continue en 0 avec $\varphi(0) = 1$ et de type positif. Alors, φ est la fonction caractéristique d'un vecteur aléatoire.*

Démonstration : Admis (pour l'instant). □

On remarque bien que ce théorème est fondamental pour la caractérisation des distributions de probabilités.

Théorème 2.5.3 (Inversion de Fourier)**.** *Soit X un vecteur aléatoire à valeurs dans $\mathbb{R}^d$, de fonction caractéristique φ_X. On suppose que φ_X est intégrable sur $\mathbb{R}^d$. Alors X admet une densité f_X continue bornée :*

$$f_X(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-itx} \varphi_X(t) dt$$

Démonstration : Comme plus haut, si h est continue à support compact

$$\mathbb{E}(h(X)) = \lim_{\lambda \to 0} \frac{1}{\pi} \int_{\mathbb{R}} h(v) \mathbb{E}\big(e^{-ivU} \varphi_X(U)\big) \frac{dv}{\lambda}$$

On a

$$\frac{1}{\lambda\pi} \mathbb{E}\big(e^{ivU} \varphi_X(U)\big) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-ivu} \varphi_X(u) e^{-\lambda|u|} du$$

Comme $e^{-\lambda|u|} \leq 1$ et $\varphi_X \in L^1$, par convergence dominée

$$\lim_{\lambda \to 0} \frac{1}{\lambda\pi} \mathbb{E}\big(e^{ivU} \varphi_X(U)\big) \to \frac{1}{2\pi} \int_{\mathbb{R}} e^{-ivu} \varphi_X(u) du$$

Par convergence dominée à nouveau

$$\mathbb{E}(h(X)) = \int_{\mathbb{R}} h(v) \left(\frac{1}{2\pi} \int_{\mathbb{R}} e^{-ivu} \varphi_X(u) du \right) dv$$

i.e. $\mathbb{P}_X$ est à densité, de densité

$$f_X(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-ivu} \varphi_X(u) du$$

□

Théorème 2.5.4 (Mutuelle indépendance et fonction caractéristique)**.** *Soit $X = (X_1, ..., X_n)$ un vecteur aléatoire. Alors les variables aléatoires (X_i) sont mutuellements indépendantes si et seulement si*

$$\varphi_X(t) = \varphi_{(X_1,...,X_n)}(t_1...t_n) = \prod_{k=1}^{n} \varphi_{X_k}(t_k), \ \forall t \in \mathbb{R}^n$$

Démonstration : On a vu que les (X_i) sont indépendants si et seulement $\mathbb{P}_X = \mathbb{P}_{(X_1,...,X_n)} = \mathbb{P}_{X_1} \otimes ... \otimes \mathbb{P}_{X_n}$.
Ainsi, comme les fonctions caratéristiques caractérisent la loi :

$$(X_i) \text{ indépendants } \iff \varphi_{\mathbb{P}_X} = \varphi_{\mathbb{P}_{X_1} \otimes ... \otimes \mathbb{P}_{X_n}}$$

Or,

$$\varphi_{\mathbb{P}_{X_1} \otimes ... \otimes \mathbb{P}_{X_n}}(t_1...t_n) = \int_{\mathbb{R}^n} e^{it_1x_1 + ... + it_nx_n} \mathbb{P}_{X_1} \otimes ... \otimes \mathbb{P}_{X_n}(dx_1...dx_n)$$

D'où,

$$\varphi_{\mathbb{P}_{X_1} \otimes ... \otimes \mathbb{P}_{X_n}}(t_1...t_n) = \int_{\mathbb{R}^n} e^{it_1x_1 + ... + it_nx_n} \mathbb{P}_{X_1}(dx_1)...\mathbb{P}_{X_n}()dx_n)$$

$$= \left(\int_{\mathbb{R}} e^{itx_1} \mathbb{P}_{X_1}(dx_1) \right) ... \left(\int_{\mathbb{R}} e^{itx_n} \mathbb{P}_{X_n}(dx_n) \right) = \varphi_{X_1}(t_1)...\varphi_{X_n}(t_n)$$

□

Proposition 2.5.2. *Si X et Y sont des variables aléatoires indépendantes alors pour tout $t \in \mathbb{R}$,*

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$$

En particulier si (X_i) indépendants et de même loi

$$\varphi_{X_1+\ldots+X_n}(t) = \varphi_{X_1}(t)^n$$

Démonstration : Si $t \in \mathbb{R}$,

$$\varphi_{X+Y}(t) = \mathbb{E}\big(e^{it(X+Y)}\big) = \mathbb{E}\big(e^{itX}e^{itY}\big) =^{\perp} \mathbb{E}\big(e^{itX}\big)\mathbb{E}\big(e^{itY}\big) = \varphi_X(t)\varphi_Y(t)$$

□

Proposition 2.5.3. *Si X un variable aléatoire de fonction caractéristique φ_X.*

- *Si $\mathbb{E}(|X|^p) < \infty$ alors $\varphi_X \in \mathcal{C}^k$, $\forall k \leq p$ et $\varphi_X^{(k)}(0) = i^k\mathbb{E}(X^k)$;*
- *Réciproquement, si φ_X est p-fois dérivable en zéro alors $\mathbb{E}(|X|^p) < \infty$, $\forall k \leq 2\left\lfloor \frac{p}{2} \right\rfloor$.*

Démonstration : Si $\mathbb{E}(|X|^p) < \infty$ alors $\mathbb{E}(|X|^k) < \infty$ pour $k \leq p$. Alors pour $\omega \in \Omega$, $\phi(t) = e^{itX(\omega)}$ est $\mathcal{C}^\infty$ et $\phi^{(k)}(t) = (iX(\omega))^k e^{itX(\omega)}$ et $|\phi^k(t)| \leq |X(\omega)|^k \in L^1$ car $\mathbb{E}(|X|^k) < \infty$.
Par dérivation sous le signe somme :
$\varphi_X(t) = \mathbb{E}(\phi(t))$ est $\mathcal{C}^k$ avec $\forall k \leq p$

$$\varphi_X^{(k)}(0) = i^k\mathbb{E}(X^k)$$

Réciproquement, si φ_X est p-fois dérivable en zéro. Montrons que X est L^{2k} pour $k \leq \left\lfloor \frac{p}{2} \right\rfloor$ par récurrence
Initialisation : Pour $k = 0$, c'est évident
Hérédité : Supposons que pour un entier $k < \left\lfloor \frac{p}{2} \right\rfloor$ on ait $X \in L^{2k}$. Alors d'après ci-dessus φ_X est $2k$-fois dérivable

$$\varphi_X^{(2k)}(t) = (-1)^k\mathbb{E}(X^{2k}e^{itX})$$

Comme $k < \left\lfloor \frac{p}{2} \right\rfloor$, $k+1 \leq \left\lfloor \frac{p}{2} \right\rfloor$ et $2k+2 \leq 2\left\lfloor \frac{p}{2} \right\rfloor \leq p$.
Ainsi, si φ_X est p-fois dérivable a fortiori φ_X est $2k+2$ fois dérivable

$$\begin{aligned}
\varphi_X^{(2k+2)}(0) &= \lim_{\epsilon\to 0} \frac{\varphi_X^{(2k)}(\epsilon) + \varphi_X^{(2k)}(-\epsilon) - 2\varphi_X^{(2k)}(0)}{\epsilon^2} \\
&= (-1)^k \lim_{\epsilon\to 0} \mathbb{E}\left(\frac{e^{i\epsilon X} + e^{-i\epsilon X} - 2}{\epsilon^2} X^{2k}\right) \\
&= (-1)^{k+1} \lim_{\epsilon\to 0} \mathbb{E}\left(\frac{\sin^2(\frac{\epsilon X}{2})}{(\frac{\epsilon X}{2})^2} X^{2k+2}\right)
\end{aligned}$$

Par Fatou,

$$\mathbb{E}(X^{2k+2}) = \mathbb{E}\left(\liminf_{\epsilon \to 0} \frac{\sin^2(\frac{\epsilon X}{2})}{(\frac{\epsilon X}{2})^2} X^{2k+2} \right) \leq \liminf \mathbb{E}\left(\frac{\sin^2(\frac{\epsilon X}{2})}{(\frac{\epsilon X}{2})^2} X^{2k+2} \right)$$

$$= (-1)^{k+1} \varphi_X^{(2k+2)}(0) < \infty$$

□

On rappelle que si $\mathbb{E}(|X|^k) < \infty$ alors φ_X est $\mathcal{C}^k$ avec

$$\varphi_X^{(k)}(t) = i^k \mathbb{E}(X^k e^{itX})$$

avec une "presque réciproque".

Exemple 2.12. Si $X \sim \mathcal{E}(\lambda)$,

$$f_X(x) = \lambda e^{-\lambda x}, \text{ sur } \mathbb{R}^+$$

$$\varphi_X(t) = \int_0^\infty e^{itx} \lambda e^{-\lambda x} dx = \frac{\lambda}{\lambda - it}$$

$$\varphi_X^{'}(t) \frac{\lambda i}{(\lambda - it)^2} \to \mathbb{E}(X) = \frac{1}{\lambda}$$

$$\varphi_X^{''}(t) = \frac{-2\lambda}{(\lambda - it)^3} \to \mathbb{E}(X^2) = \frac{2}{\lambda^2}$$

Et on retrouve bien $\mathrm{Var}(X) = \frac{1}{\lambda^2}$.

En résumé et comme pour toutes les autres fonctions introduites ci-dessus, la fonction caractéristique permet de caractériser entièrement une distribution de probabilité et est souvent utilisée pour simplifier les calculs et les preuves.

On se rappellera par exemple que la fonction caractéristique détermine de manière unique la distribution de probabilité d'une variable aléatoire.
Les fonctions caractéristiques sont aussi particulièrement utiles pour les calculs impliquant des sommes de variables aléatoires indépendantes. Par exemple, si X et Y sont des variables aléatoires indépendantes alors la fonction caractéristique de la somme $X + Y$ est simplement le produit des fonctions caractéristiques,

$$\varphi_{X+Y}(t) = \mathbb{E}(e^{itX} e^{itY}) =^{\perp} \mathbb{E}(e^{itX}) \mathbb{E}(e^{itY}) = \varphi_X(t) \varphi_Y(t)$$

On voit aussi dans l'expression

$$\varphi_X(t) = \mathbb{E}(e^{itX})$$

un lien évident avec les moments de X. En effet, on a vu que les dérivées de la fonction caractéristique en $t = 0$ donnent les moments de la distribution :

$$\mathbb{E}(X) = \frac{1}{i} \varphi_X^{'}(0), \ \mathbb{E}(X^2) = \frac{1}{i^2} \varphi_X^{''}(0)$$

Pour aller plus loin que la simple somme de variables aléatoires, la fonction est plus généralement utile pour étudier des transformations de variables aléatoires en raison de ses propriétés algébriques. Par exemple, considérons la translation $Y = X + a$, alors

$$\varphi_Y(t) = \mathbb{E}(e^{it(X+a)}) = e^{ita}\mathbb{E}(e^{itX}) = e^{ita}\varphi_X(t)$$

Considérons un autre exemple, la multiplication (changement d'échelle). Soit $Z = bX$, alors

$$\varphi_Z(t) = \mathbb{E}(e^{itbX}) = \varphi_X(bt)$$

De ces deux exemples, on peut déduire la transformation affine. Soit $\Lambda = aX + b$, alors

$$\varphi_\Lambda(t) = e^{itb}\varphi_X(at)$$

On verra plus tard que la fonction caractéristique joue un rôle crucial dans l'étude des théorèmes limites comme le théorème central limite.

2.5.2 Autres transformées exponentielles

Définition 2.5.2 (Fonction génératrice)**.** Soit X une variable aléatoire à valeurs dans $\mathbb{N}$. On appelle fonction génératrice de X la fonction $G_X : [0,1] \to \mathbb{R}$ et

$$G_X(s) = \mathbb{E}(s^X) = \sum_{k\geq 0} s^k \mathbb{P}(X = k)$$

Plus généralement, si $X = (X_1, ..., X_d)$ à valeurs dans $\mathbb{N}^d$, on pose

$$G_X(s_1, ..., s_d) = \mathbb{E}(s_1^{X_1} ... s_d^{X_d})$$

On remarque première le lien formel entre la fonction génératrice et la fonction caractéristique

$$G_X(e^{it}) = \varphi_X(t)$$

De plus, on remarque que comme la fonction génératrice est définie par une série entière dont le rayon de convergence est ≥ 1, la fonction $s \mapsto G_X(s)$ est $\mathcal{C}^\infty$ sur $[0,1[$ et

$$G_X^{(n)}(0) = n!\mathbb{P}(X = n)$$

Ce qui implique bien évidemment que la fonction génératrice caractérise la loi (distribution).

Finalement, on remarque que par le théorème de dérivation sous le signe somme, G_X est n-fois dérivable en 1 si et seulement si $\mathbb{E}(X^n) < \infty$ et

$$G_X^{(n)}(1) = \mathbb{E}(X(X-1)...(X-n+1))$$

On note l'apparition d'un moment dit factoriel.

Exemple 2.13. Voyons l'exemple de calcul de fonction génératrice classique.
Supposons que $X \sim \mathcal{B}(p)$, alors

$$G_X(s) = (1-p) + ps$$

Si maintenant $Y \sim \mathcal{G}(p)$, alors

$$G_Y(s) = \frac{ps}{1-(1-p)s}$$

Puis finalement si $Z \sim \mathcal{P}(\lambda)$, alors

$$G_Z(s) = e^{\lambda(s-1)}$$

Proposition 2.5.4. *Soit X et Y à valeurs dans $\mathbb{N}$ alors*

1. $X \perp Y \iff G_{(X,Y)}(s,t) = G_X(s)G_Y(t)$
2. $X \perp Y \implies G_{X+Y}(t) = G_X(t)G_Y(t)$

Exemple 2.14. Si $X \sim \mathcal{B}(n,p)$, on a déjà vu l'expression de la fonction génératrice d'une variable aléatoire suivant une loi de Bernoulli. On en déduit ainsi que

$$G_X(s) = ((1-p) + ps)^n$$

Maintenant, si $Y \sim \mathcal{P}(\lambda)$ et $Z \sim \mathcal{P}(\mu)$, alors comme les lois sont indépendantes,

$$G_{X+Y}(s) = G_X(s)G_Y(s) = e^{(\lambda+\mu)(s-1)}$$

D'où $X + Y \sim \mathcal{P}(\lambda + \mu)$ car la fonction génératrice caractérise la loi.

Passons maintenant à une autre fonction intéressante : La transformée de Laplace.

Définition 2.5.3 (Transformée de Laplace)**.** Soit X un vecteur aléatoire à valeurs dans $\mathbb{R}^d$. On définit sa transformée de Laplace $L_X : \mathbb{R}^d \to [0, +\infty]$ par la formule

$$L_X(t) = \mathbb{E}(e^{t.X})$$

où $t = (t_1, ..., t_d)$ et $t.X = \sum_{i=1}^{d} t_i X_i$ est le produit scalaire usuel sur cet espace.

On remarque encore une fois un lien formel avec la fonction caractéristique

$$L_X(it) = \varphi_X(t)$$

Et on remarque de plus que, contrairement à la fonction caractéristique ou la fonction génératrice, L_X peut valoir $+\infty$.

Proposition 2.5.5. *Soit X une variable aléatoire*

1. $L_X(0) = 1$;

2. *Si $L_X(t) < \infty$ sur un voisinage de 0 alors elle est développable en série entière :*

$$L_X(t) = \sum_{n\geq 0} \frac{t^n}{n!}\mathbb{E}(X^n) \text{ et } \mathbb{E}(X^n) = L_X^{(n)}(0)$$

Si elle est finie sur un voisinage V de 0, la transformée de Laplace caractérise la loi.
En effet, si tel est le cas alors la fonction de la variable complexe $z \mapsto H_X(z) = \mathbb{E}(e^{zX})$ est bien définie sur l'ouvert $\mathcal{O} = \{z = x + iy \mid x \in V,\ y \in \mathbb{R}\}$ et elle y est holomorphe.
Restreinte à $i\mathbb{R}$, H_X n'est autre que la transformée de Fourier de X.
Le théorème de prolongement analytique dit que si deux fonctions holomorphes coïncident sur un ensemble qui possède un point d'accumulation, alors elles coïncident sur leur ouvert maximal de définition.
Ici, H_X coïncident avec L_X sur V et avec φ_X sur $i\mathbb{R}$. Ainsi si on connait L_X sur V, par prolongement analytique , on connait φ_X et comme φ_X caractérise la loi...

On remarque de plus que si X est une variable positive, on pourra considérer alternativement

$$\tilde{L}_X(t) = \mathbb{E}(e^{-tX}) \text{ avec } t \geq 0$$

Proposition 2.5.6. *Si X et Y sont des variables aléatoires réelles,*

1. $X \perp Y \iff L_{(X,Y)}(s,t) = L_X(s)L_Y(t)$
2. $X \perp Y \implies L_{X+Y}(t) = L_X(t)L_Y(t)$.

Exemple 2.15. Soit $X \sim \mathcal{E}(\lambda)$, alors

$$f_X(x) = \lambda e^{-\lambda x}\mathbb{1}_{x\geq 0}$$

Puis,

$$\tilde{L}_X(t) = \mathbb{E}(e^{-tX}) = \int_0^\infty e^{-tx}\lambda e^{-\lambda x}dx = \frac{\lambda}{\lambda + t}$$

Maintenant, on considère $Y \sim \Gamma(n,\lambda)$, alors

$$f_Y(x) = \frac{\lambda^n}{\Gamma(n)}x^{n-1}e^{-\lambda x}\mathbb{1}_{x\geq 0}$$

D'où,

$$\tilde{L}_Y(t) = \mathbb{E}(e^{-tY}) = \frac{\lambda^n}{\Gamma(n)}\int_0^\infty x^{n-1}e^{-(t+\lambda)x}dx = \frac{\lambda^n}{\Gamma(n)}\frac{\Gamma(n)}{(\lambda+\mu)^n} = \frac{\lambda^n}{(\lambda+\mu)^n}$$

On voit alors que si $X_1, ..., X_n \sim \mathcal{E}(\lambda)$ indépendantes mutuellements alors $X_1 + ... + X_n \sim \Gamma(n,\lambda)$.

2.6 Probabilités, lois et espérances conditionelles

Dans cette section, nous allons surtout rappeler des acquis de prépa.

2.6.1 Probabilités conditionelles

Définition 2.6.1 (Probabilité conditionnelle)**.** Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé et $A \in \mathcal{F}$ avec $\mathbb{P}(A) > 0$. On définit la probabilité conditionnelle sachant A comme $\mathbb{P}(.|A) : \mathcal{F} \to [0,1]$ et

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}, \ \forall B \in \mathcal{F}$$

On remarque, comme à chaque fois que l'on définit un objet nommé "probabilité", que c'est bien une probabilité.

$$\mathbb{P}\left(\bigsqcup_i B_i | A\right) = \frac{\mathbb{P}\left(\bigsqcup_i B_i \cap A\right)}{\mathbb{P}(A)} = \sum_i \frac{\mathbb{P}(B_i \cap A)}{\mathbb{P}(A)} = \sum_i \mathbb{P}(B_i|A)$$

De plus, on remarque que

$$A \perp B \Longrightarrow \mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} =^{\perp} \frac{\mathbb{P}(B)\mathbb{P}(A)}{\mathbb{P}(A)} = \mathbb{P}(B)$$

Théorème 2.6.1 (Formule des probabilités totales)**.** *Soit (A_i) des ensembles dénombrables tels que $\Omega = \bigsqcup_i A_i$ (partitions de l'univers) avec $\mathbb{P}(A_i) > 0$. Alors pour tout $B \in \mathcal{F}$ on a*

$$\mathbb{P}(B) = \sum_i \mathbb{P}(B|A_i)\mathbb{P}(A_i)$$

Démonstration : On écrit simplement $B = B \cap \Omega = \bigsqcup_i (B \cap A_i)$. □

Exemple 2.16. $\Omega = A \sqcup \overline{A}$ avec $\mathbb{P}(A) \in]0,1[$

$$\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|\overline{A})\mathbb{P}(\overline{A})$$

Théorème 2.6.2 (Formule de Bayes)**.** *Si $\Omega = \bigsqcup_i A_i$ avec $\mathbb{P}(A_i) > 0$ et si $\mathbb{P}(B) > 0$ alors*

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)}$$

Démonstration : On utilise deux fois la formule des probabilités conditionnelles

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)}$$

et on conclut avec la formule des probabilités totales au dénominateur. □

On voit maintenant l'exemple classique que tout élève voyant ce cours devrait connaître

Exemple 2.17. Un virus atteint $0,5\%$ de la population. Un test T permet de décider si on est malade ou pas avec la fiabilité suivante :

1. $T\oplus$ pour 95% des malades.
2. $T\ominus$ pour 99% des non-malades.

On note M pour malade. Alors d'une part

$$\mathbb{P}(M) = \frac{0,5}{100} \text{ et } \mathbb{P}(\overline{M}) = \frac{99,5}{100}$$

Maintenant calculons

$$\mathbb{P}(M|\oplus) = \frac{\mathbb{P}(\oplus|M)\mathbb{P}(M)}{\mathbb{P}(\oplus|M)\mathbb{P}(M) + \mathbb{P}(\oplus|\overline{M})\mathbb{P}(\overline{M})} = \frac{\frac{95}{100}\frac{0,5}{100}}{\frac{95}{100}\frac{0,5}{100} + \frac{1}{100}\frac{99,5}{100}} \equiv 0,32$$

2.6.2 Loi et espérance conditionelle

Définition 2.6.2 (Loi et espérance conditionnelle)**.** Soit X une variable aléatoire discrète et $x \in X(\Omega)$ tel que $\mathbb{P}(X = x) > 0$. La loi conditionnelle sachant $\{X = x\}$ est définit par :

$$\mathbb{P}^{X=x}(B) = \mathbb{P}(B|X = x),\ \forall B \in \mathcal{F}$$

On appelle espérance conditionnelle sachant $\{X = x\}$ et on note $\mathbb{E}(.|X = x)$ l'espérance sous la loi $\mathbb{P}^{X=x}$.

Exemple 2.18. On lance 2 dés. On note X et Y les résultats.

$$\mathbb{E}(X + Y|X + Y = 8) = \sum_{k=1}^{6} k\mathbb{P}(X = k|X + Y = 8) = \sum_{k=1}^{6} k\frac{\mathbb{P}(X = k \text{ et } X + Y = 8)}{\mathbb{P}(X + Y = 8)}$$

D'où,

$$\mathbb{E}(X + Y|X + Y = 8) = \sum_{k=1}^{6} k\frac{\mathbb{P}(X = k, Y = 8 - k)}{\mathbb{P}(X + Y = 8)} = \sum_{k=1}^{6} k\frac{\mathbb{P}(X = k)\mathbb{P}(Y = 8 - k)}{\mathbb{P}(X + Y = 8)}$$

Or, si $k = 1$ alors on a $\mathbb{P}(Y = 8 - k) = \mathbb{P}(Y = 7) = 0$, donc

$$\mathbb{E}(X + Y|X + Y = 8) = \sum_{k=2}^{6} k\frac{\mathbb{P}(X = k)\mathbb{P}(Y = 8 - k)}{\mathbb{P}(X + Y = 8)} = \sum_{k=2}^{6} k\frac{\frac{1}{3}6}{\frac{5}{3}6} = 4$$

Par symétrie,

$$\mathbb{E}(Y|X + Y = 8) = \mathbb{E}(X|X + Y = 8)$$

Et, $\mathbb{E}(X + Y|X + Y = 8) = 8$ donc

$$\mathbb{E}(Y|X + Y) = 4$$

Définition 2.6.3 (Densité conditionnelle)**.** Soit (X,Y) un couple de variable aléatoire à densité $f_{(X,Y)}(x,y)$. On définit la densité conditionnelle de Y sachant $\{X=x\}$

$$f_{Y|X=x}(y) = \frac{f_{(X,Y)}(x,y)}{f_X(x)} = \frac{f_{(X,Y)}(x,y)}{\int f_{(X,Y)}(x,y)dy}$$

C'est une densité de probabilité.
L'espérance associée est l'espérance conditionnelle sachant $\{X=x\}$, i.e. si h est mesurable bornée

$$\mathbb{E}(h(Y)|X=x) = \int h(y) f_{Y|X=x}(y)dy$$

Exemple 2.19. Soit (X,Y) de densité $f_{(X,Y)}(x,y) = \frac{1}{x}e^{-\frac{y}{x}-x}$ avec $x,y \geq 0$.
Alors

$$\int f_{(X,Y)}(x,y)dy = e^{-x}\int_0^\infty \frac{1}{x}e^{-\frac{y}{x}}dy = e^{-x}$$

Cela donne donc

$$f_{Y|X=x}(y) = \frac{1}{x}e^{-\frac{y}{x}}\mathbb{1}_{y\geq 0}$$

C'est-à-dire,

$$\mathcal{L}(Y|X=x) = \mathcal{E}(\frac{1}{x})$$
$$\mathbb{E}(Y|X=x) = x$$

2.7 Convergence des variables aléatoires

On commence par rappeler le lemme de Borel-Cantelli.

Lemme 2.7.1 (Borel-Cantelli)**.** *Soit $(\Omega,\mathcal{F},\mathbb{P})$ un espace probabilisé et (A_n) une suite d'évènements. Alors,*

1. *Si $\sum_{n\geq 1}\mathbb{P}(A_n) < \infty$ alors*
$$\mathbb{P}(\limsup A_n) = 0$$
2. *Si les (A_n) sont mutuellements indépendants alors si $\sum_{n\geq 1}\mathbb{P}(A_n) = \infty$, alors*
$$\mathbb{P}(\limsup A_n) = 1$$

<u>*Démonstration :*</u> On rappelle que

$$\limsup A_n = \bigcap_{n\geq 0}\bigcup_{k\geq n} A_k$$

avec $\bigcup_{k\geq n} A_k$ décroissant en n. Ainsi,

$$\mathbb{P}(\limsup A_n) = \lim \mathbb{P}(\bigcup_{k\geq n} A_k)$$

Si bien que,

(a) Si $\sum \mathbb{P}(A_n) < \infty$ alors par le lemme de Boole

$$\mathbb{P}(\bigcup_{k\geq n} A_k) \leq \sum_{k\geq n} \mathbb{P}(A_k) \to 0$$

Si bien que

$$\mathbb{P}(\limsup A_n) = 0$$

(b) Supposons maintenant les (A_n) mutuellement indépendants

$$\mathbb{P}(\overline{\limsup A_n}) = \mathbb{P}(\liminf \overline{A_n}) = \mathbb{P}(\bigcup_{n\geq 0} \bigcap_{k\geq n} \overline{A_k})$$

avec $\bigcap_{k\geq n} \overline{A_k}$ croissant en n.

$$\mathbb{P}(\overline{\limsup A_n}) = \lim \mathbb{P}(\bigcap_{k\geq n} \overline{A_k})$$

Par indépendance des (A_n)

$$\mathbb{P}(\bigcap_{k\geq n} \overline{A_k}) = \prod_{k\geq n} \mathbb{P}(\overline{A_k}) = \prod_{k\geq n} (1 - \mathbb{P}(A_k))$$

et,

$$\log \mathbb{P}(\bigcap_{k\geq n} \overline{A_k}) = \sum_{k\geq n} \log(1 - \mathbb{P}(A_k))$$

Or, par concavité du log

$$\log(1-x) \leq -x$$

D'où,

$$\log \mathbb{P}(\bigcap_{k\geq n} \overline{A_k}) = \sum_{k\geq n} \log(1 - \mathbb{P}(A_k)) \leq \sum_{k\geq n} -\mathbb{P}(A_k) = -\sum_{k\geq n} \mathbb{P}(A_k) = -\infty$$

Ainsi, $\mathbb{P}(\bigcap_{k\geq n} \overline{A_k}) = 0$ et par suite $\mathbb{P}(\overline{\limsup A_n}) = 0$. C'est-à-dire

$$\mathbb{P}(\limsup A_n) = 1$$

□

2.7.1 Modes de convergence : Convergence presque sûre

Définition 2.7.1 (Convergence presque sûre)**.** On dit qu'une suite (X_n) de variables aléatoires converge presque sûrement vers une variable aléatoire X et on note $X_n \to^{p.s.} X$ si

$$\mathbb{P}(\{\omega \in \Omega \mid \lim X_n(\omega) = X(\omega)\}) = 1$$

On remarque que

$$X_n \to^{p.s.} X \iff \forall p \geq 1 \ \mathbb{P}(\bigcup_{n\geq 1} \bigcap_{k_g eqn} \{|X_k - X| < \frac{1}{p}\}) = 1$$

$$\iff \forall \epsilon > 0 \ \mathbb{P}(\bigcup_{n\geq 1} \bigcap_{k\geq n} |X_k - X| < \epsilon) = 1$$

$$\iff \forall \epsilon > 0 \ \mathbb{P}(\liminf\{|X_n - X| < \epsilon\}) = 1$$

$$\iff \forall \epsilon > 0 \ \mathbb{P}(\limsup\{|X_n - X| > \epsilon\}) = 0$$

Corollaire 2.7.1. *Soient (X_n) et X des variables aléatoires.*

1. *Si pour tout $\epsilon > 0$, $\sum_{n\geq 1} \mathbb{P}(|X_n - X| > \epsilon) < \infty$ alors X_n convergence presque sûrement vers X ;*
2. *Si les (X_n) sont mutuellement indépendants alors*

$$X_n \to^{p.s.} 0 \iff \sum_{n\geq 1} \mathbb{P}(|X_n| > \epsilon) < \infty$$

Démonstration : On utilise la définition de la convergence simple et le lemme de Borel-Cantelli qui fait directement tombé le résultat. □

Proposition 2.7.1. *Si $X_n \to^{p.s.} X$ et si f est continue alors*

$$f(X_n) \to^{p.s.} f(X)$$

Démonstration : Par continuité de f

$$f(X_n(\omega)) \to f(X)$$

Dès lors que $X_n(\omega) \to X(\omega)$

$$\{\omega \mid X_n(\omega) \to X(\omega)\} \subseteq \{\omega \mid f(X_n(\omega)) \to f(X(\omega))\}$$

D'où,

$$1 = \mathbb{P}(\{\omega \mid X_n(\omega) \to X(\omega)\}) \leq \mathbb{P}(\{\omega \mid f(X_n(\omega)) \to f(X(\omega))\})$$

C'est-à-dire, $\mathbb{P}(\{\omega \mid f(X_n(\omega)) \to f(X(\omega))\}) = 1$. □

A ce stade il est important de noter que la convergence presque sûre est une forme de convergence très forte. Elle implique que la convergence se produit pour chaque ω dans l'espace probabilisé, ce qui est plus restrictif que d'autres types de convergence que nous verrons par la suite.

2.7.2 Modes de convergence : Convergence en probabilité

Définition 2.7.2 (Convergence en probabilité)**.** On dit qu'une suite (X_n) de variables aléatoires converge en probabilité vers une variable aléatoire X et on note $X_n \to^{\mathbb{P}} X$ si

$$\forall \epsilon > 0, \lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0$$

Cela signifie donc que la probabilité que X_n soit proche de X devient de plus en plus grande à mesure que n augmente. Cette condition est plus faible que la convergence simple. Nous verrons cela plus en détail dans les sections d'articulations entre chaque modes de convergence.

Proposition 2.7.2 (Unicité de la limite)**.** *Si $X_n \to^{\mathbb{P}} X$ et $X_n \to^{\mathbb{P}} Y$ alors $X = Y$ presque sûrement.*

Démonstration : Pour $p \geq 1$ on remarque que

$$\{|X - Y| > \frac{1}{p}\} \subseteq \{|X_n - X| > \frac{1}{2p}\} \cup \{|X_n - Y| > \frac{1}{2p}\}$$

De sorte que,

$$\mathbb{P}(|X - Y| > \frac{1}{p}) \leq \mathbb{P}(|X_n - X| > \frac{1}{p}) + \mathbb{P}(|X_n - Y| > \frac{1}{p})$$

Et par convergence en probabilité

$$\mathbb{P}(|X_n - X| > \frac{1}{p}), \mathbb{P}(|X_n - Y| > \frac{1}{p}) \to 0$$

C'est-à-dire,

$$\mathbb{P}(|X - Y| > \frac{1}{p}) = 0$$

Si bien que,

$$\mathbb{P}(X \neq Y) = \mathbb{P}(\bigcup_{p\geq 1} |X - Y| > \frac{1}{p}) = 0$$

□

Proposition 2.7.3. *Si $X_n \to^{\mathbb{P}} X$ et f continue alors $f(X_n) \to^{\mathbb{P}} f(X)$.*

Démonstration : f est uniformément continue sur les compacts.
Soit $\epsilon > 0$ et $a > 0$, s'il existe $\eta > 0$ tq $|x| \leq a$, $|x - y| < \eta$ alors $|f(x) - f(y)| \leq \epsilon$.
En passant au complémentaire,

$$\{|f(X_n) - f(X)| \geq \epsilon\} \subseteq \{|X| > a\} \cup \{|X_n - X| > \eta\}$$

Donc,

$$\mathbb{P}(|f(X_n) - f(X)| \geq \epsilon) \leq \mathbb{P}(|X| > a) + \mathbb{P}(|X_n - X| > \eta)$$

$$\limsup_n \mathbb{P}(|f(X_n) - f(X)| \geq \epsilon) \leq \mathbb{P}(|X| > a) + \limsup_n \mathbb{P}(|X_n - X| > \eta)$$

La dernière lim sup valant 0

$$\limsup_n \mathbb{P}(|f(X_n) - f(X)| \geq \epsilon) \leq \mathbb{P}(|X| > a)$$

En faisant $a \to \infty$, $\mathbb{P}(|X| > a) \to 0$.

$$\lim \mathbb{P}(|f(X_n) - f(X)| \geq \epsilon) = 0$$

□

On remarque que si on pose $d_{\mathbb{P}}(X, Y) = \mathbb{E}(\min(|X-Y|, 1))$ alors $X_n \to^{\mathbb{P}} X \iff d_{\mathbb{P}}(X_n, X) \to 0$. Autrement dit, $d_{\mathbb{P}}$ métrise la convergence en probabilité. En fait l'espace $(L^0(\Omega, \mathcal{F}, \mathbb{P}), d_{\mathbb{P}})$ est un espace métrique complet.

2.7.3 Modes de convergence : Convergence L^p

Définition 2.7.3 (Convergence L^p)**.** Soit $p \geq 1$. On dit (X_n) converge vers X dans L^p et $X_n \to^{L^p} X$ si $||X_n - X||_p \to_n 0$, c'est-à-dire $\mathbb{E}(|X_n - X|^p) \to_n 0$.

On remarque que par l'inégalité de Hölder, si X_n converge vers X dans L^p alors X_n converge dans X dans L^q pour $q \leq p$.
Par exemple,

$$\mathbb{E}(|X_n - X) \leq \mathbb{E}(|X_n - X|^2)^{\frac{1}{2}}$$

2.7.4 Modes de convergence : Convergence en loi

Définition 2.7.4 (Convergence en loi)**.** On dit que (X_n) converge en loi vers X et on note $X_n \to^{\mathcal{L}} X$ si pour tout f continue bornée

$$\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$$

La convergence ci-dessus s'écrit pour tout f continue bornée

$$\int f(x)\mathbb{P}_{X_n}(dx) \to \int f(x)\mathbb{P}_X(dx)$$

C'est la convergence étroite de $\mathbb{P}_{X_n}$ vers $\mathbb{P}_X$.
La convergence en loi ne concerne que les lois de X_n et X.

Lemme 2.7.2 (Porte-manteau)**.** *On a l'équivalence entre*

- *X_n converge vers X en loi ;*
- *Pour tout F fermé,* $\limsup \mathbb{P}(X_n \in F) \leq \mathbb{P}(X \in F)$ *;*
- *Pour tout O ouvert,* $\liminf \mathbb{P}(X_n \in O) \geq \mathbb{P}(X \in O)$ *;*
- *Pour tout B borélien tel que $\mathbb{P}(X \in \overline{B} \backslash \mathring{B}) = 0$, on a* $\lim \mathbb{P}(X_n \in B) = \mathbb{P}(X \in B)$.

Démonstration : (1) $\Rightarrow$ (2) : Si F est fermé on considère

$$f_k(x) = \frac{1}{(1+d(x,F))^k}$$

f_k est continue bornée et décroit vers $\mathbb{1}_F$.

$$\limsup \mathbb{P}(X_n \in F) = \limsup \int \mathbb{1}_F(x)\mathbb{P}_{X_n}(dx) \leq \limsup \int f_k(x)\mathbb{P}_{X_n}(dx)$$

Or d'une part

$$\limsup \int f_k(x)\mathbb{P}_{X_n}(dx) = \limsup \mathbb{E}(f_k(X_n))$$

Et par convergence en loi

$$\limsup \mathbb{E}(f_k(X_n)) = \mathbb{E}(f_k(X)) \to \mathbb{E}(\mathbb{1}_F(X)) = \mathbb{P}(X \in F)$$

C'est-à-dire,

$$\limsup \mathbb{P}(X_n \in F) \leq \mathbb{P}(X \in F)$$

(2) $\Rightarrow$ (3) : On passe au complémentaire

(2) + (3) $\Rightarrow$ (4) : On a $\mathring{B} \subseteq B \subseteq \overline{B}$ de suite

$$\mathbb{P}(X \in \mathring{B}) \leq \liminf \mathbb{P}(X_n \in \mathring{B}) \leq \liminf \mathbb{P}(X_n \in B) \leq \limsup \mathbb{P}(X_n \in B)$$

$$\leq \limsup \mathbb{P}(X_n \in \overline{B}) \leq \mathbb{P}(X \in \overline{B})$$

Si $\mathbb{P}(X \in \overline{B}\backslash\mathring{B}) = 0$ alors on a

$$\liminf \mathbb{P}(X_n \in B) = \limsup \mathbb{P}(X_n \in B) = \mathbb{P}(X \in B)$$

(4) $\Rightarrow$ (1) : Soit f continue bornée, $a \leq f \leq b$. Quitte à considérer $\tilde{f} = \frac{f-a}{b-a}$, on peut se ramener à $0 \leq f \leq 1$.

$$\mathbb{E}(f(X_n)) = \int_0^1 \mathbb{P}(f(X_n) > t)dt$$

$$\mathbb{E}(f(X)) = \int_0^1 \mathbb{P}(f(X) > t)dt$$

$$\mathbb{P}(f(X) > t) = \mathbb{P}(X \in f^{-1}(]t, +\infty[) = B)$$

On a $\overline{B}\backslash\mathring{B} = \overline{f^{-1}(]t,+\infty[)}\backslash f^{-1}(]t,+\infty[)$.
$\mathbb{P}(X \in \overline{B}\backslash\mathring{B}) \leq \mathbb{P}(X = t) = 0$ sauf en un ensemble D_0 au plus dénombrable de t.
$\mathbb{P}(X_n \in \overline{B}\backslash\mathring{B}) \leq \mathbb{P}(X_n = t) = 0$ sauf en un ensemble au plus dénombrable D_n.
L'ensemble $\bigcup_{n\geq 0} D_n$ est au plus dénombrable et si $t \notin \bigcup_{n\geq 0} D_n$,

$$\mathbb{P}(f(X_n) > t) \to \mathbb{P}(f(X) > t)$$

Par convergence dominée dans les expressions de $\mathbb{E}(f(X_n))$ et $\mathbb{E}(f(X))$ on conclut

$$\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$$

□

Théorème 2.7.1 (Lien convergence en loi / fonction de répartition)**.** *Soit (X_n) une suite de variables aléatoires réelles, alors $X_n \to^{\mathcal{L}} X$ si et seulement si $F_{X_n}(t) \to F_X(t)$ pour tout t tels que F_X est continue en t.*

Démonstration : $\Rightarrow$: Soit $t \in \mathbb{R}$ et $B =]-\infty, t]$. Si F_X est continue en t alors

$$\mathbb{P}(X \in \overline{B} \backslash \mathring{B}) = \mathbb{P}(X = t) = 0$$

Si X_n converge en loi vers X alors

$$F_{X_n}(t) = \mathbb{P}(X_n \in B) \to \mathbb{P}(X \in B) = F_X(t)$$

$\Leftarrow$: Réciproquement, supposons que $F_{X_n}(t) \to F_X(t)$ en tous les points de continuité de F_X. Prenons $f \in \mathcal{C}_c^\infty$ (support compact), alors

$$f(x) = \int_{-\infty} f'(t)dt = \int f'(t)\mathbb{1}_{]-\infty,x]}(t)dt$$

Par Fubini

$$\mathbb{E}(f(X)) = \int f'(t)\mathbb{E}(\mathbb{1}_{]-\infty,X]}(t))dt = \int f'(t)\mathbb{E}(\mathbb{1}_{X \geq t})dt$$

$$= \int f'(t)\mathbb{P}(X \geq t)dt = \int f'(t)(1 - F_X(t^-))dt$$

De même,

$$\mathbb{E}(f(X_n)) = \int f'(t)(1 - F_{X_n}(t))dt$$

Comme l'ensemble des points de discontinuité de F_X est au plus dénombrable, on a pour Lebesgue presque tout t

$$f'(t)(1 - F_{X_n}(t^-)) \to f'(t)(1 - F_X(t^-))$$

Par convergence dominée, on déduit

$$\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$$

On conclut par densité de $\mathcal{C}_c^\infty$ dans $\mathcal{C}_b$. □

Théorème 2.7.2 (Lien convergence en loi / fonction caractéristique)**.** *La suite (X_n) converge vers X en loi si et seulement si $\forall t \in \mathbb{R}$, $\varphi_{X_n}(t) \to \varphi_X(t)$.*

Démonstration : $\Leftarrow$: On traite d'abord le sens réciproque, soit $h \in \mathcal{C}_c^2$ (support compact).

$$\hat{h}(t) = \int e^{itx} h(x)dx \text{ est intégrable}$$

En effet, en l'infini on a alors

$$|\hat{h}(t)| = \mathcal{O}\left(\frac{1}{t^2}\right)$$

car

$$\hat{h}(t) = \int_{\mathbb{R}} e^{itx} h(x)dx = \int \frac{e^{itx}}{it} h'(x)dx = \frac{1}{t^2} \int e^{itx} h''(x)dx$$

Par inversion de Fourier, on a alors

$$h(x) = \frac{1}{2\pi} \int e^{-itx} \hat{h}(t)dt$$

Par Fubini, on a alors

$$\mathbb{E}(h(X_n)) = \frac{1}{2\pi} \int \mathbb{E}(e^{itX_n})\hat{h}(t)dt = \frac{1}{2\pi} \int \overline{\varphi_{X_n}(t)}\hat{h}(t)dt$$

Par convergence dominée, on a alors :

$$\mathbb{E}(h(X_n)) \to \mathbb{E}(h(X))$$

On conclut par densité de $\mathcal{C}_c^2$ dans $\mathcal{C}_b$.

$\Rightarrow$: Le sens direct est immédiat car $x \mapsto e^{itx}$ est continue.

□

Théorème 2.7.3 (de Lévy)**.** *Soit (φ_{X_n}) une suite de fonctions caractéristique qui converge ponctuellement vers φ, i.e. $\forall t \in \mathbb{R},\ \varphi_{X_n}(t) \to \varphi(t)$.*
Si φ est continue en 0, alors φ est une fonction caractéristique $\varphi = \varphi_X$ par une variable aléatoire X et alors X_n converge vers X en loi.

Démonstration : Admis (pour le moment). □

Théorème 2.7.4 (Continuous mapping theorem)**.** *Soit $X_n \to^{\mathcal{L}} X$ et f continue ($\mathbb{P}_X$ p.s.) alors $f(X_n) \to^{\mathcal{L}} f(X)$.*

Démonstration : Si f est continue et h est $\mathcal{C}_b$ alors $h \circ f$ est continue bornée de sorte que

$$\mathbb{E}(h \circ f(X_n)) \to \mathbb{E}(h \circ f(X)) \text{ i.e. } f(X_n) \to^{\mathcal{L}} f(X)$$

Si f est continue $\mathbb{P}_X$ p.s., on utilise le lemme du Portemanteau. □

2.7.5 Articulation des modes de convergence : Convergences p.s. et en probabilité

Lemme 2.7.3. *Si $X_n \to^{p.s.} X$ alors $X_n \to^{\mathbb{P}} X$.*

Démonstration : On a que vu que $X_n \to^{p.s.} X$ si et seulement si

$$\forall \epsilon > 0,\ \mathbb{P}(\limsup |X_n - X| > \epsilon) = 0$$

C'est-à-dire,

$$\forall \epsilon > 0,\ \mathbb{P}\Big(\bigcap_{n\geq 0} \bigcup_{k\geq n} |X_k - X| > \epsilon \Big) = 0$$

Alors, comme $\bigcup_{k\geq n} |X_k - X| > \epsilon$ est décroissant en n, on a

$$\forall \epsilon > 0,\ \lim \mathbb{P}\Big(\bigcup_{k\geq n} |X_k - X| > \epsilon \Big) = 0$$

D'où,

$$\forall \epsilon > 0,\ \lim \mathbb{P}\Big(|X_n - X| > \epsilon \Big) = 0$$

C'est-à-dire,

$$X_n \to^{\mathbb{P}} X$$

□

Lemme 2.7.4. *Si $X_n \to^{\mathbb{P}} X$ alors il existe une suite strictement croissante (n_k) telle que*

$$X_{n_k} \to^{p.s.} X$$

Démonstration : Si $X_n \to^{P} X$, pour $k \geq 1$ soit n_k le plus petit entier tel que

$$\mathbb{P}\Big(|X_{n_k} - X| > \frac{1}{k} \Big) \leq \frac{1}{2^k}$$

Par le lemme de Borel-Cantelli :

$$\mathbb{P}\Big(\limsup |X_{n_k} - X| > \frac{1}{k} \Big) = 0$$

i.e. pour $\mathbb{P}$ presque tout $\omega \in \Omega$, il existe $k_0(\omega)$ tel que si $k \geq k_0(\omega)$ alors

$$|X_{n_k}(\omega) - X(\omega)| < \frac{1}{k}$$

C'est-à-dire

$$X_{n_k} \to^{p.s.} X$$

□

2.7.6 Articulation des modes de convergence : Convergences L^p, p.s., $\mathbb{P}$

Lemme 2.7.5. *Si $X_n \to^{p.s.} X$ et $|X_n| \leq Y \in L^1$ alors $X_n \to^{L^1} X$.*

Démonstration : C'est la convergence dominée. □

Lemme 2.7.6. *Si $X_n \to^{L^p} X$ alors $X_n \to^{\mathbb{P}} X$.*

Démonstration : Soit $\epsilon > 0$ par l'inégalité de Markov

$$\mathbb{P}\bigg(|X_n - X| > \epsilon\bigg) \leq \frac{\mathbb{E}(|X_n - X|^p)}{\epsilon^p}$$

□

Définition 2.7.5. On dit qu'une famille de variables aléatoires $(X_i)_{i\in I}$ est uniformément intégrable (équi-intégrable) si

$$\lim_{n\to+\infty} \sup_{i\in I} \mathbb{E}(|X_i|\mathbb{1}_{|X_i|>n}) = 0$$

Exemple 2.20. Si $\forall i, |X_i| \leq Y \in L^1$ alors

$$\sup_{i\in I} \mathbb{E}\bigg(|X_i|\mathbb{1}_{|X_i|>n}\bigg) \leq \mathbb{E}\bigg(|Y|\mathbb{1}_{|Y|>n}\bigg) \to 0$$

Exemple 2.21. Si (X_i) est bornée dans L^p avec $p \geq 1$ alors (X_i) est uniformément intégrable.

$$\text{Bornée dans } L^p \iff \sup_{i\in I} \mathbb{E}(|X_i|^p) < \infty$$

En effet, par l'inégalité de Hölder, si $\dfrac{1}{p} + \dfrac{1}{q} = 1$

$$\mathbb{E}(|X_i|\mathbb{1}_{|X_i|>n}) \leq \mathbb{E}(|X_i|^p)^{\frac{1}{p}} \mathbb{P}(|X_i| > n)^{\frac{1}{q}}$$

Par l'inégalité de Markov,

$$\leq \mathbb{E}(|X_i|^p)^{\frac{1}{p}} \left(\frac{\mathbb{E}(|X_i|)}{n}\right)^{\frac{1}{q}}$$

$$\sup_{i\in I} \mathbb{E}(|X_i|\mathbb{1}_{|X_i|>n}) \leq \left(\sup_{i\in I} \mathbb{E}(|X_i|^p)\right)^{\frac{1}{p}} \left(\sup_{i\in I} \mathbb{E}(|X_i|)\right)^{\frac{1}{q}} \frac{1}{n^{\frac{1}{q}}} \to_{n\to 0} 0$$

Proposition 2.7.4. *La famille (X_i) est uniformément intégrable si et seulement*

- *(X_i) est bornée dans L^1, $\sup_{i\in I} \mathbb{E}(|X_i|) < \infty$;*
- *$\forall \epsilon > 0,\ \exists \delta > 0$ si $A \in \mathcal{F}$, $\mathbb{P}(A) < \delta$ alors $\sup_{i\in I} \mathbb{E}(|X_i|\mathbb{1}_A) < \epsilon$*

Démonstration : Plus tard. □

Théorème 2.7.5. *Soit (X_n) une suite de variables aléatoires intégrables. Alors il y a équivalence entre*

1. $X_n \to^{L^1} X$;
2. (X_n) *est uniformément intégrable et* $X_n \to^{\mathbb{P}} X$.

Démonstration : (1) $\Rightarrow$ (2) : On a vu $X_n \to^{L^1} X \Rightarrow X_n \to^{\mathbb{P}} X$.
Il reste à voir que (X_n) est uniformément intégrable.
Soit $\epsilon > 0$, $\exists n_0 >> 1$ tel que $\forall n \geq n_0$

$$\mathbb{E}(|X_n - X|) \leq \frac{\epsilon}{2}$$

Soit $A \in \mathcal{F}$,

$$\mathbb{E}(|X_n|\mathbb{1}_A) \leq \mathbb{E}(|X_n - X|\mathbb{1}_A) + \mathbb{E}(|X|\mathbb{1}_A) \leq \frac{\epsilon}{2} + \mathbb{E}(|X|\mathbb{1}_A)$$

Or $\{X\}$ est uniformément intégrable donc $\exists \delta > 0$ tel que si $\mathbb{P}(A) < \delta$ alors $\mathbb{E}(|X|\mathbb{1}_A) < \frac{\epsilon}{2}$, donc

$$\forall n \geq n_0,\ \mathbb{E}(|X_n|\mathbb{1}_A) \leq \epsilon$$

C'est-à-dire, (X_n) est uniformément intégrable.

(2) $\Rightarrow$ (1) : Comme $X_n \to^{\mathbb{P}} X$, il existe une sous-suite n_k tel que $X_{n_k} \to^{p.s.} X$

$$\mathbb{E}(|X|) = \mathbb{E}(\liminf_k |X_{n_k}|) \leq \liminf_k \mathbb{E}(|X_{n_k}|) \leq \sup_n \mathbb{E}(|X_n|) < \infty$$

car uniformément intégrable. Donc $X \in L^1$.
Alors pour $\epsilon > 0$

$$\mathbb{E}(|X_n - X|) = \mathbb{E}(|X_n - X|\mathbb{1}_{|X_n - X| > \epsilon}) + \mathbb{E}(|X_n - X|\mathbb{1}_{|X_n - X| \leq \epsilon})$$

et $\mathbb{E}(|X_n - X|\mathbb{1}_{|X_n - X| \leq \epsilon}) \leq \epsilon$ d'où

$$\mathbb{E}(|X_n - X|) \leq \mathbb{E}(|X_n|\mathbb{1}_{|X_n - X| > \epsilon}) + \mathbb{E}(|X|\mathbb{1}_{|X_n - X| > \epsilon}) + \epsilon$$

Comme (X_n) est uniformément intégrable et $X \in L^1$ alors $((X_n) \cup \{X\})$ est aussi uniformément intégrable et par ailleurs $X_n \to^{\mathbb{P}} X$ donc

$$\mathbb{P}(|X_n - X| > \epsilon) \to 0$$

Pour n assez grand, on a

$$\mathbb{E}(|X_n|\mathbb{1}_{|X_n - X| > \epsilon}) < \epsilon$$

$$\mathbb{E}(|X|\mathbb{1}_{|X_n - X| > \epsilon}) < \epsilon$$

Si bien que

$$\mathbb{E}(|X_n - X|) \leq 3\epsilon$$

□

2.7.7 Articulation des modes de convergences : Convergence en loi et autres modes

Lemme 2.7.7. *Si $X_n \to^{\mathbb{P}} X$ alors $X_n \to^{\mathcal{L}} X$.*

Démonstration : On suppose que X_n converge en probabilité vers X, i.e.

$$\forall \epsilon > 0,\ \mathbb{P}(|X_n - X| > \epsilon) \to 0$$

Soit $t \in \mathbb{R}$ et soit $\epsilon << 1$.

$$|\varphi_{X_n}(t) - \varphi_X(t)| = |\mathbb{E}(e^{itX_n}) - \mathbb{E}(e^{itX})| = |\mathbb{E}(e^{itX_n} - e^{itX})|$$

Pour progresser, on doit utiliser une inégalité classique sur la différence de deux exponentielle.

$$|e^{itX_n} - e^{itX}| = |e^{itX}||e^{itX_n}e^{-itX} - 1| = |e^{it(X_n - X)} - 1|$$

On développe en série de Taylor

$$e^{it(X_n - X)} = 1 + it(X_n - X) + \mathcal{O}((t(X_n - X))^2)$$

D'où

$$|e^{it(X_n - X)} - 1| \sim |it(X_n - X)| = |t||X_n - X|$$

D'autre part, par périodicité et par le fait que l'exponentielle soit bornée

$$|e^{it(X_n - X)} - 1| \leq 2$$

Si bien que

$$|e^{itX_n} - e^{itX}| \leq |t| \min\{|X_n - X|, 2\}$$

On va utiliser cette inégalité.

$$|\varphi_{X_n}(t) - \varphi_X(t)| = |\mathbb{E}(e^{itX_n} - e^{itX})| \leq |t|\mathbb{E}(\min(|X_n - X|, 2))$$
$$\leq |t|\mathbb{E}(|X_n - X|\mathbb{1}_{|X_n - X| < \epsilon}) + |t|\mathbb{E}(\min(|X_n - X|, 2)\mathbb{1}_{|X_n - X| \geq 1})$$
$$\leq |t|\epsilon + 2|t|\mathbb{P}(|X_n - X| > \epsilon)$$

Si bien qu'en passant à la lim inf

$$\liminf |\varphi_{X_n}(t) - \varphi_X(t)| \leq |t|\epsilon$$

On fait ensuite $\epsilon \to 0$ et on en déduit la convergence en loi.
□

Corollaire 2.7.2. *Si $X_n \to^{p.s., L^1} X$ alors $X_n \to^{\mathcal{L}} X$.*

Lemme 2.7.8. *Si $X_n \to^{\mathcal{L}} c \in \mathbb{R}$ alors $X_n \to^{\mathbb{P}} c$.*

Démonstration : Soit $\epsilon > 0$

$$\mathbb{P}(|X_n - c| > \epsilon) = \mathbb{P}(X_n < c - \epsilon) + \mathbb{P}(X_n > c + \epsilon)$$

La variable aléatoire c a pour fonction de répartition

$$F_c(t) = 0 \text{ si } t < c \text{ ou } 1 \text{ si } t \geq c$$

Elle est continue sur $]-\infty, c[$ et $[c, +\infty[$.

$$\mathbb{P}(|X_n - c| > \epsilon) = \mathbb{P}(X_n < c - \epsilon) + \mathbb{P}(X_n > c + \epsilon)$$

Par convergence en loi

$$\mathbb{P}(X_n < c - \epsilon) \to F_c((c-\epsilon)^-) = 0$$

$$\mathbb{P}(X_n > c + \epsilon) \to 1 - F_c(c + \epsilon) = 1 - 1 = 0$$

Si bien que

$$\mathbb{P}(|X_n - c| > \epsilon) \to 0$$

□

2.7.8 Convergence des variables aléatoires : Résumé

- $X_n \to^{p.s.} X$ ssi $\mathbb{P}(\{\omega \in \Omega \mid X_n(\omega) \to X(\omega)\}) = 1$.
- $X_n \to^{\mathbb{P}} X$ ssi $\forall \epsilon > 0,\ \mathbb{P}(|X_n - X| > \epsilon) \to 0$.
- $X_n \to^{L^p} X$ ssi $\mathbb{E}(|X_n - X|^p) \to 0$.
- $X_n \to^{\mathcal{L}} X$ ssi $\forall f \in \mathcal{C}_b,\ \mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$ ssi $\forall t \in \mathbb{R},\ \varphi_{X_n}(t) \to \varphi_X(t)$ ssi $\forall t$ où F_X est continue $F_{X_n}(t) \to F_X(t)$.

$$\begin{array}{ccccc} \text{Convergence } L^p & \Longrightarrow & \text{Convergence en } \mathbb{P} & \Longleftarrow & \text{Convergence p.s.} \\ & & \Downarrow & & \\ & & \text{Convergence en } \mathcal{L} & & \end{array}$$

2.8 Théorèmes limites

Dans tout le chapitre, on se place sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ sur lequel est définie une suite de variables $(X_n)_{n \geq 0}$, indépendantes mutuellement et identiquement distribuées (i.i.d.).

2.8.1 Loi des grands nombres (LGN) : Loi faible des grands nombres

Avant d'énoncer et montrer la loi faible des grands nombres, on va démontrer un lemme clé de la démonstration.

Lemme 2.8.1. *Si $(a_i), (b_i) \in \mathbb{C}^{\mathbb{N}}$ avec $|a_i| \leq 1$, $|b_i| \leq 1$ alors*

$$\left| \prod_{i=1}^{n} a_i - \prod_{i=1}^{n} b_i \right| \leq \sum_{i=1}^{n} |b_i - a_i|$$

Démonstration : On démontre le lemme par récurrence sur n.
Si on suppose le prédicat vrai pour n fixé

$$\left|\prod_{i=1}^{n+1} a_i - \prod_{i=1}^{n+1} b_i\right| \leq |a_{n+1}|\left|\prod_{i=1}^{n} a_i - \prod_{i=1}^{n} b_i\right| + |a_{n+1} - b_{n+1}|\left|\prod_{i=1}^{n} b_i\right|$$

avec, $|a_{n+1}| \leq 1$ et $\left|\prod_{i=1}^{n} b_i\right| \leq 1$. Par hypothèse de récurrence

$$\leq \sum_{i=1}^{n} |a_i - b_i| + |a_{n+1} - b_{n+1}| = \sum_{i=1}^{n+1} |a_i - b_i|$$

□

Théorème 2.8.1 (Loi faible des grands nombres)**.** *Soit (X_n) une suite de variables aléatoires indépendantes et identiquement distribuées intégrables, i.e. $\mathbb{E}(|X_1|) < \infty$, alors*

$$\frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \to^{\mathbb{P}} \mathbb{E}(X_1)$$

Démonstration : On suppose d'abord que $X_i \in L^2$. Soit $\epsilon > 0$, on remarque que

$$\mathbb{E}\left(\frac{S_n}{n}\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}n\mathbb{E}(X_1) = \mathbb{E}(X_1)$$

Par l'inégalité de Tchebychev

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mathbb{E}(X_1)\right| > \epsilon\right) \leq \frac{\mathrm{Var}(\frac{S_n}{n})}{\epsilon^2} = \frac{\mathrm{Var}(S_n)}{n^2\epsilon^2} = \frac{n\mathrm{Var}(X_1)}{n^2\epsilon^2} = \frac{\mathrm{Var}(X_1)}{n\epsilon^2} \to 0$$

Maintenant, si $X_i \in L^1$ et pas L^2, on utilise le lemme précédent.
Quitte à considérer $X_i - \mathbb{E}(X_i)$ au lieu de X_i, on peut se ramener au cas où $\mathbb{E}(X_i) = 0$ et montrer que $\frac{S_n}{n} \to^{\mathbb{P}} 0$ et donc on aurait $\frac{S_n}{n} \to^{\mathcal{L}} 0$.
Soit $t \in \mathbb{R}$, par indépendance

$$\varphi_{\frac{S_n}{n}}(t) = \mathbb{E}\left(e^{i\frac{t}{n}\sum X_i}\right) = \varphi_{X_1}^n\left(\frac{t}{n}\right)$$

D'après le lemme précédent

$$\left|\varphi_{\frac{S_n}{n}}(t) - 1\right| \leq n\left|\varphi_{X_1}\left(\frac{t}{n}\right) - 1\right|$$

Comme $\mathbb{E}(|X_1|) < \infty$, φ_{X_1} est dérivable en zéro avec

$$\varphi_{X_1}^{'}(0) = i\mathbb{E}(X_1) = 0$$

On écrit alors (en reconnaissant un taux d'accroissement en 0)

$$t.\frac{n}{t}\left|\varphi_{X_1}\left(\frac{t}{n}\right)-1\right| \to t|\varphi_{X_1}^{'}(0)| = 0$$

En conclusion, on a pour tout $t \in \mathbb{R}$

$$\left|\varphi_{\frac{S_n}{n}}(t)-1\right| \to 0$$

i.e. $\frac{S_n}{n} \to^{\mathcal{L}} 0$ et par suite

$$\frac{S_n}{n} \to^{\mathbb{P}} 0$$

□

Ce théorème est fondamental en théorie des probabilités car il décrit le comportement de la moyenne d'un grand nombre de variables aléatoires i.i.d.

La moyenne empirique $\frac{1}{n}\sum_{i=1}^{n} X_i$ converge en probabilité vers la moyenne théorique :

$$\forall \epsilon > 0,\ \lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X_1)\right| > \epsilon\right) = 0$$

Cela signifie que si on prend un grand nombre de variables aléatoires i.i.d. et que l'on calcule leur moyenne, cette moyenne empirique sera de plus en plus proche de l'espérance mathématique à mesure que le nombre de variables augmente.

De part l'interprétation qualitative, on peut espérer des applications assez importantes du théorème :

- Statistiques : La loi faible des grands nombres est utilisée pour justifier l'utilisation de la moyenne empirique comme estimateur de la moyenne théorique d'une population.
- Simulations : En simulation, elle garantit que la moyenne des résultats de plusieurs simulations convergera vers la moyenne théorique.
- Assurance et finance : Elle est utilisée pour estimer les risques et les rendements moyens sur de grands ensembles de données.

On remarque aussi une chose, c'est que la moyenne empirique divisée par n relève de l'aléatoire tandis que l'espérance de la variable aléatoire X_1 relève du déterminisme. On en parlera plus après avoir parlé de la loi forte des grands nombres.

2.8.2 Loi des grands nombres (LGN) : Loi forte des grands nombres

Maintenant, voyons un théorème bien similaire à la loi faible des grands nombres mais avec une conclusion plus forte.

Théorème 2.8.2 (Loi forte des grands nombres)**.** *Soit (X_n) une suite de variables aléatoire indépendantes et identiquement distribuées intégrables, i.e. $\mathbb{E}(|X_1|) < \infty$, alors*

$$\frac{S_n}{n} = \frac{1}{n}\sum_{k=1}^{n} X_i \to^{p.s.,L^1} \mathbb{E}(X_1)$$

Démonstration : Plus tard. □

Remarque 2.1.

- Si (X_n) est i.i.d. et vérifie

$$(*) \quad \frac{S_n}{n} \to^{p.s.} c \in \mathbb{R}$$

alors, $\mathbb{E}(|X_1|) < \infty$ et $\mathbb{E}(X_1) = c$.
En effet, si $(*)$ alors

$$\frac{X_n}{n} = \frac{S_n - S_{n-1}}{n} = \frac{S_n}{n} - \frac{S_{n-1}}{n-1}\frac{n-1}{n} \to^{p.s.} 0$$

par Borel-Cantelli, on déduit que

$$\forall \epsilon > 0, \ \sum \mathbb{P}\left(\left|\frac{X_n}{n}\right| > \epsilon\right) < \infty$$

En particulier,

$$\sum \mathbb{P}\left(\frac{|X_n|}{n} > 1\right) < \infty$$

Et,

$$\sum \mathbb{P}\left(\frac{|X_n|}{n} > 1\right) = \sum \mathbb{P}\left(|X_n| > n\right) = \sum \mathbb{P}\left(|X_1| > n\right) = \mathbb{E}(|X_1|) < \infty$$

- Si X_i i.i.d. positives avec $\mathbb{E}(X_1) = +\infty$ alors

$$\frac{S_n}{n} \to^{p.s.} \mathbb{E}(X_1) = +\infty$$

La loi des grands nombres a une portée profonde : "Du hasard peut naître le déterminisme". Elle légitime également la méthode scientifique, c'est-à-dire répéter plusieurs fois une expérience dans des conditions identiques pour prévoir un résultat théorique.

2.8.3 Applications de la LGN

Méthode de Monte-Carlo :

Cette méthode est une technique de simulation utilisée pour estimer des valeurs numériques, souvent dans des contextes où des solutions analytiques sont difficiles ou impossibles à obtenir. Elle repose sur l'utilisation de nombres aléatoires pour effectuer des simulations répétées et estimer des quantités d'intérêts.

Le principe est le suivant. La méthode commence par générer un grand nombre de nombres aléatoires selon une certaine distribution. Puis, ces nombres aléatoires sont utilisés pour simuler le processus ou le phénomène d'intérêt. Enfin, les résultats des simulations sont utilisés pour estimer la quantité d'intérêt, souvent en calculant une moyenne ou autre.

Mathématiquement,

- On génère $X_1, ..., X_n$ indépendantes et identiquement distribuées selon $\mathbb{P}_X$.
- On calcule la moyenne empirique
$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
- On utilise $\overline{X_n}$ comme estimateur de $\mathbb{E}(X)$.

L'exemple classique de l'utilisation de la méthode de Monte-Carlo est l'estimation de la valeur de π. C'est lent, pas optimisé mais la méthode est facile à comprendre. Voilà comment cela fonctionne :

- On part de l'hypothèse que nous disposons d'un cercle de rayon r inscrit dans un carré de côtés $2r$.
- Maintenant, on génère un grand nombres de points aléatoires uniformément distribués dans ce carré et on se pose la question : Quelle est la probabilité qu'un point aléatoire soit à l'intérieur du cercle ?
$$\mathbb{P}(\text{Point dans le cercle}) = \frac{\mathbb{P}(\text{Dans le cercle})}{\mathbb{P}(\text{Dans le carré})} = \frac{\text{Aire du cercle}}{\text{Aire du carré}} = \frac{\pi r^2}{4r^2} = \frac{\pi}{4}$$
- On multiplie cette proportion de points dans le cercle par 4 et on obtient une estimation de π.

Maintenant, passons au vrai sujet.

La loi des grands nombres joue un rôle central dans la méthode de Monte-Carlo en garantissant que la moyenne des résultats des simulations converge vers la valeur théorique attendue à mesure que le nombre de simulations augmente.

Un autre exemple classique est l'estimation de l'intégrale.

Si (X_i) sont indépendants et identiquement distribués de loi $\mathbb{P}_X$ et si $g : \mathbb{R} \to \mathbb{R}$ tel que $g(X_i) \in L^1$ alors

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \to^{p.s.,L^1} \mathbb{E}(g(X_1)) = \int g(x)\mathbb{P}_X(dx)$$

Par exemple, si $\mathbb{P}_X$ est à densité f_X

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \to^{p.s.} \int g(x) f_X(x) dx$$

Estimation paramétrique :

Vote avec 2 issues possibles A ou B. Le résultat final est une proportion p pour A et $1-p$ pour B.
On choisit "au hasard" un échantillon "représentatif" des votants de taille n.

$$X_i = 1 \text{ si } A$$

$$X_i = 0 \text{ si } B$$

Si l'échantillon est "vraiment" représentatif, on doit avoir $\mathbb{P}(X_i = 1) = p$ et $\mathbb{P}(X_i = 0) = 1 - p$, c'est-à-dire X_i i.i.d. $\mathcal{B}(p)$.
D'après la LGN

$$\frac{S_n}{n} \equiv \mathbb{E}(X_1) = p$$

La question légitime est alors, à quelle vitesse

$$\frac{S_n}{n} \to p \ ?$$

2.8.4 Théorème Central Limite (TCL)

Le Théorème central limite est un rafinement de la Loi des grands nombres.
Avant de passer à l'énoncé et à la preuve de ce théorème, on aura besoin d'un lemme :

Lemme 2.8.2. *Pour tout $x \in \mathbb{R}$, $p \geq 1$*

$$\left| e^{ix} - \sum_{k=0}^{p} \frac{(ix)^k}{k!} \right| \leq \min\left(\frac{|x|^{p+1}}{(p+1)!}, \frac{2|x|^p}{p!} \right)$$

Démonstration : On applique la formule de Taylor avec reste intégral :

$$e^{ix} - \sum_{k=0}^{p} \frac{(ix)^k}{k!} = \frac{i^{p+1}}{p!} \int_0^x (x-s)^p e^{is} ds$$

On a

$$\left| e^{ix} - \sum_{k=0}^{p} \frac{(ix)^k}{k!} \right| \leq \frac{1}{p!} \left| \int_0^x (x-s)^p ds \right| = \frac{|x|^{p+1}}{(p+1)!}$$

Par ailleurs, par une intégration par parties donne

$$\left|e^{ix} - \sum_{k=0}^{p} \frac{(ix)^k}{k!}\right| \leq \left|\frac{i^p}{(p+1)!}\int_0^x (x-s)^{p-1} e^{is} ds - \frac{(ix)^p}{p!}\right| \leq \frac{|x|^p}{p!} + \frac{|x|^p}{p!} = \frac{2|x|^p}{p!}$$

□

Théorème 2.8.3 (Central Limite). *Soit (X_n) une suite de variables aléatoires indépendantes et identiquement distribuées avec $\mathbb{E}(|X_i|^2) < \infty$. On pose $\sigma^2 = Var(X_1)$. Alors*

$$\sqrt{n}\left(\frac{S_n}{n} - \mathbb{E}(X_1)\right) \to^{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

de façon équivalente

$$\frac{\sqrt{n}}{\sigma}\left(\frac{S_n}{n} - \mathbb{E}(X_1)\right) \to^{L} \mathcal{N}(0, 1)$$

Démonstration : Quitte à considérer $\dfrac{X_i - \mathbb{E}(X_i)}{\sigma}$, on peut supposer que $\mathbb{E}(X_i) = 0$ et $\mathrm{Var}(X_i) = 1$. On calcule la fonction caractéristique, par indépendance, pour $t \in \mathbb{R}$

$$\varphi_{\frac{S_n}{\sqrt{n}}}(t) = \varphi_{X_1}(\frac{t}{\sqrt{n}})^n$$

Comme $\mathbb{E}(X^2) < \infty$, φ_{X_1} est 2 fois dérivable et

$$\varphi_{X_1}(t) = 1 + t\varphi_{X_1}^{'}(0) + \frac{t^2}{2}\varphi_{X_1}^{''}(0) + o(t^2) = 1 + 0 - \frac{t^2}{2} + o(t^2)$$

D'après le lemme, en prenant $a_i = \varphi_{X_1}\left(\dfrac{t}{\sqrt{n}}\right)$, $b_i = \left(1 - \dfrac{t^2}{2n}\right)$, on a

$$\left|\varphi_{\frac{S_n}{\sqrt{n}}}(t) - \left(1 - \frac{t^2}{2n}\right)^n\right| \leq n\left|\varphi_{X_1}\left(\frac{t}{\sqrt{n}}\right) - \left(1 - \frac{t^2}{2n}\right)\right|$$

D'après le lemme d'avant pour $p = 2$,

$$\left|\varphi_{X_1}\left(\frac{t}{\sqrt{n}}\right) - \left(1 - \frac{t^2}{2n}\right)\right| \leq \mathbb{E}\left(\min\left(\frac{|t|^3|X_1|^3}{6n^{\frac{3}{2}}}, \frac{t^2|X_1|^2}{n}\right)\right)$$

Ainsi par convergence dominée

$$\left|\varphi_{\frac{S_n}{\sqrt{n}}}(t) - \left(1 - \frac{t^2}{2n}\right)^n\right| \leq \mathbb{E}\left(\min\left(\frac{|t|^3|X_1|^3}{6\sqrt{n}}, t^2|X_1|^2\right)\right) \to 0$$

Par ailleurs,

$$\left(1 - \frac{t^2}{2n}\right)^n = \exp\left(n\log\left(1 - \frac{t^2}{2n}\right)\right) \to \exp\left(-\frac{t^2}{2}\right)$$

Par conséquent,

$$\left|\varphi_{\frac{S_n}{\sqrt{n}}}(t) - \exp\left(-\frac{t^2}{2}\right)\right| = \left|\varphi_{\frac{S_n}{\sqrt{n}}}(t) - \varphi_{\mathcal{N}(0,1)}(t)\right| \to 0$$

□

La Théorème Central Limite assure enfait que la vitesse de convergence dans la Loi des Grands Nombres est $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$.
La loi gaussienne $\mathcal{N}(0,1)$ apparaît comme une loi limite universelle.

On remarque que le Théorème Central Limite peut s'énoncer, de manière équivalente, de la manière suivante :

$$\forall [a,b] \subseteq \mathbb{R},\ \mathbb{P}\left(\frac{\sqrt{n}}{\sigma}\left(\frac{S_n}{n} - \mathbb{E}(X_1)\right) \in [a,b]\right) \to \frac{1}{2\pi}\int_a^b \exp\left(-\frac{x^2}{2}\right)dx$$

Théorème 2.8.4. *Soit (X_n) une suite de variables aléatoires indépendantes avec $Var(X_i) < \infty, \forall i \in \mathbb{N}^*$. On pose $\sigma_n = \sqrt{Var(X_1) + ... + Var(X_n)}$. On suppose de plus la condition de Lindeberg :*

$$\forall \epsilon > 0,\ \frac{1}{\sigma_n^2}\sum_{i=1}^n \mathbb{E}\left(\left|X_i - \mathbb{E}(X_i)\right|^2 \mathbb{1}_{|X_i - \mathbb{E}(X_i)| > \epsilon\sigma_n}\right) \to 0 \quad (*)$$

Alors,

$$Z_n = \frac{1}{\sigma_n}\sum_{i=1}^n (X_i - \mathbb{E}(X_i)) \to^{\mathcal{L}} \mathcal{N}(0,1)$$

Démonstration : Comme plus haut, par indépendance

$$\forall t \in \mathbb{R},\ \varphi_{Z_n}(t) = \prod_{k=1}^n \mathbb{E}\left(e^{i\frac{t}{\theta_n}(X_k - \mathbb{E}(X_k))}\right)$$

Avec,

$$\mathbb{E}\left(e^{i\frac{t}{\theta_n}(X_k - \mathbb{E}(X_k))}\right) = 1 - \frac{t^2}{2\theta_n^2}\mathrm{Var}(X_k) + o\left(\frac{t^2}{\theta_n^2}\right)$$

D'après le lemme sur la différence de 2 produits, on a alors

$$\left|\varphi_{Z_n}(t) - \prod_{k=1}^n (1 - \frac{t^2}{2\theta_n^2}\mathrm{Var}(X_k))\right| \leq \frac{1}{\sigma_n^2}\sum_{i=1}^n \mathbb{E}\left(\left|X_i - \mathbb{E}(X_i)\right|^2 \mathbb{1}_{|X_i - \mathbb{E}(X_i)| > \epsilon\sigma_n}\right) \to 0$$

On remarque enfin que

$$\prod_{k=1}^n (1 - \frac{t^2}{2\theta_n^2}\mathrm{Var}(X_k)) = \exp\left(\sum_{k=1}^n \log\left(1 - \frac{t^2}{2\theta_n^2}\mathrm{Var}(X_k)\right)\right) = \exp\left(-\frac{t^2}{2}.1 + o(1)\right)$$

□

2.8.5 Retour sur les applications de la LGN

Méthode de Monte-Carlo :

On a vu que si X_i indépendantes et identiquement distribuées et si $g : \mathbb{R} \to \mathbb{R}$ telle que $g(X_i) \in L^1$ alors

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \to \mathbb{E}(g(X_1)) = \int g(x)\mathbb{P}_X(dx)$$

A l'aide du TCL, on peut préciser

$$\frac{\sqrt{n}}{\mathrm{Var}(g(X_1))}\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}(g(X_1))\right) \to \mathcal{N}(0,1)$$

Si bien que la vitesse de convergence de $\displaystyle\sum_{i=1}^{n} g(X_i)$ vers $\mathbb{E}(g(X_1))$ est de l'ordre de $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$.
Cette vitesse est indépendantes de la régularité de la fonction g, contrairement aux autres méthodes d'intégrations usuelles.

Estimation paramétrique :

On reprend un exemple précédent. $X_i \sim \mathcal{B}(p)$ (p inconnue) indépendantes et identiquement distribuées avec $X_i = 1$ si votre pour A et $X_i = 0$ si vote pour B.
On a vu que la LGN implique

$$\frac{S_n}{n} \to \mathbb{E}(X_1) = p$$

Fixons un seuil de confiance $1 - \alpha = 0,95$ par exemple.
Ensuite, il existe η_α tel que

$$\mathbb{P}(|\mathcal{N}(0,1)| \leq \eta_\alpha) \geq 1 - \alpha$$

D'après la LGN dans le TCL, on a

$$\mathbb{P}\left(\frac{\sqrt{n}}{\sigma}\left|\frac{S_n}{n} - p\right| \leq \eta_\alpha\right) \to \mathbb{P}(|\mathcal{N}(0,1)| \leq \eta_\alpha) \geq 1 - \alpha$$

Et,

$$\mathbb{P}\left(\frac{\sqrt{n}}{\sigma}\left|\frac{S_n}{n} - p\right| \leq \eta_\alpha\right) = \mathbb{P}\left(p \in \left[\frac{S_n}{n} - \frac{n_\alpha\sigma}{\sqrt{n}}, \frac{S_n}{n} + \frac{n_\alpha\sigma}{\sqrt{n}}\right]\right)$$

Comme $\sigma = p(1-p) \leq \frac{1}{4}$ et $1,96 \leq 2$, ainsi

$$\mathbb{P}\left(p \in \left[\frac{S_n}{n} - \frac{1}{2\sqrt{n}}, \frac{S_n}{n} + \frac{1}{2\sqrt{n}}\right]\right) \geq 1 - \alpha = 95\%$$

Autrement dit, avec une probabilité de 95% de ne pas se tromper on peut affirmer que

$$p \in \left[\frac{S_n}{n} - \frac{1}{2\sqrt{n}}, \frac{S_n}{n} + \frac{1}{2\sqrt{n}}\right]$$

La taille de la fourchette est $\frac{1}{\sqrt{n}}$.
Ainsi, si on interroge 10^4 personnes, on a une précision 10^{-2}.

2.8.6 Vitesse de convergence dans le TCL

Théorème 2.8.5 (Berry-Esseen). *Soit (X_i) une suite de variables aléatoires indépendantes et identiquement distribuées avec $\mathbb{E}(|X_i|^3) < \infty$. Alors,*

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{n}}{\sigma}\left(\frac{S_n}{n} - \mathbb{E}(X_1) \right) \leq x \right) - \mathbb{P}(\mathcal{N}(0,1) \leq x) \right| \leq \frac{c.\mathbb{E}(|X_1|^3)}{Var(X_1)\sqrt{n}}$$

où C est une constante universelle ≤ 3.

Démonstration : Plus tard.

2.8.7 Fonctions de répartition empirique

Dans cette section, on considère (X_i) une suite de variables aléatoires indépendantes et identiquement distribuées. D'après la LGN et le TCL, si $\mathbb{E}(X_i^2) < \infty$ on a

$$\frac{1}{n}\sum_{k=1}^{n} X_k \to^{p.s.,L^1} \mathbb{E}(X_1)$$

$$\sqrt{n}\left(\frac{1}{n}\sum_{k=1}^{n} X_k - \mathbb{E}(X_1) \right) \to^{\mathcal{L}} \mathcal{N}(0, \mathrm{Var}(X_1))$$

On s'intéresse aux fonctions de répartitions empiriques

$$F_n(t) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{X_k \leq t}$$

Si les (X_k) sont indépendantes et indentiquement dribtibuées alors les variables $\mathbb{1}_{X_k\leq t}$ le sont aussi, de loi de Bernoulli

$$\mathbb{1}_{X_i\leq t} \sim \mathcal{B}(p), \text{ avec } p = \mathbb{P}(X_1 \leq t) = F_X(t)$$

D'après la LGN, on a pour t fixé

$$F_n(t) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{X_k\leq t} \to^{p.s.,L^1} \mathbb{E}(\mathbb{1}_{X_1\leq t}) = F_X(t)$$

D'après le TCL on a de plus

$$\sqrt{n}(F_n(t) - F_X(t)) \to^{\mathcal{L}} \mathcal{N}(0, F_X(t)(1 - F_X(t)))$$

On aimerait avoir des versions fonctionnelles de ces théorèmes.

2.8.8 Fonctions de répartition empirique : Lemme de Dini

Ces lemmes permettent de passer d'une convergence simple de fonctions à une convergence uniforme sous des hypothèses de monotonie : monotonie de la suite de fonctions ou des fonctions elle-mêmes.

Lemme 2.8.3 (de Dini). *Soient $a < b$ dans $\mathbb{R}$ et $f_n : [a,b] \to \mathbb{R}$ une suite de fonctions qui converge simplement vers $f : [a,b] \to \mathbb{R}$.*

1. *Si la suite (f_n) est croissante et si les f_n et f sont continues, alors la convergence est uniforme.*
2. *Si les f_n sont des fonctions croissantes et si f est continue, alors la convergence est uniforme.*

Démonstration : 1. On pose $g_n = f - f_n \geq 0$ continue et qui, par hypothèses, converge simplement vers 0.
Il s'agit de voir que $\lim\limits_{n\infty} \sup\limits_{[a,b]} g_n(x) = 0$.
Par l'absurde, si $\exists \epsilon > 0$, n_k tel que $\sup\limits_{[a,b]} g_{n_k}(x) > \epsilon$. Alors on a

$$\sup_{[a,b]} g_{n_k}(x) = \sup_{[a,b]} g_{n_k}(x^*_{n_k})$$

avec $x^*_{n_k} \in [a,b]$.
On peut alors extraire une sous-suite qui converge vers $x_\infty \in [a,b]$.
On aura alors $\lim\limits_k \sup g_{n_k}(x^*_{n_k}) \geq \epsilon$ et $\lim\limits_k \sup g_{n_k}(x_\infty) \geq \epsilon$. Ce qui contredit la convergence simple.

2. Si f est continue sur $[a,b]$ qui est un compact alors elle est uniformément continue. Pour tout $\epsilon > 0$, il existe une subdivision $[a,b]$ comme suit

$$a_1 = a < a_2 < a_3 < ... < a_k = b$$

telle que

$$0 \leq f(a_{i+1}) - f(a_i) \leq \epsilon, \quad 1 \leq i \leq k-1$$

Pour $x \in [a,b]$, on prend i tel que $a_i \leq x \leq a_{i+1}$ par monotonie

$$f_n(x) - f(x) \leq f_n(a_{i+1}) - f(a_i) \leq f_n(a_{i+1}) - f(a_{i+1}) + \epsilon$$

$$f_n(x) - f(x) \geq f_n(a_i) - f(a_{i+1}) \geq f_n(a_i) - f(a_i) - \epsilon$$

Par convergence simple en les (a_i), il existe N_ϵ tel que $\forall n \geq N_\epsilon$, $\forall i \in \{1, ..., k\}$, $|f_n(a_i) - f(a_i)| \leq \epsilon$.
Par suite,

$$|f_n(x) - f(x)| \leq 2\epsilon, \ \forall x \in [a,b]$$

C'est-à-dire,

$$\sup_{[a,b]} |f_n(x) - f(x)| \leq 2\epsilon$$

et f_n converge uniformément vers f. □

Corollaire 2.8.1. *Soit F_n une suite de fonction de répartition qui converge simplement vers une fonction de répartition F continue. Alors, F_n converge uniformément vers F.*

Démonstration : On adapte la preuve du point 2. des lemmes de Dini. Comme F est continue et, $F(-\infty) = 0$ et $F(+\infty) = 1$, il existe une subdivision finie de $\overline{\mathbb{R}}$

$$-\infty = a_1 < a_2 < ... < a_k = +\infty$$

telle que

$$0 \leq F(a_{i+1}) - F(a_i) \leq \epsilon$$

La suite de la preuve est identique. □

2.8.9 Fonctions de répartition empirique : Théorème de Gliverko-Cantelli

Théorème 2.8.6 (de Gliverko-Cantelli)**.** *Soit (X_i) une suite de variables aléatoires indépendantes et identiquement distribuée, alors*

$$\mathbb{P}\left(\limsup_{n\infty} \sup_{t\in\mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_k \leq t} - F_X(t) \right| = 0 \right) = 1$$

i.e. presque sûrement la suite des fonctions de répartition empiriques converge vers la fonction de répartition limite.
i.e. $\mathbb{P}(\lim_{n\infty} ||F_n - F_X||_\infty = 0) = 1$.

Démonstration : On commence par se ramener au cas où les X_i sont uniformes sur $[0, 1]$. En effet, on se souvient $X_i \sim F_X^{-1}(U_i)$ avec U_i indépendantes et identiquement distribuée de loi uniforme sur $[0, 1]$. Ainsi,

$$\sup_{t\in\mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq t} - F_X(t) \right| =^{\mathcal{L}} \sup_{t\in\mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{F_X^{-1}(U_i) \leq t} - F_X(t) \right|$$

$$= \sup_{s\in F_X(\mathbb{R})} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{U_i \leq s} - s \right|, \text{ en posant} F_X(t) = s$$

$$\leq \sup_{s\in[0,1]} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{U_i \leq s} - s \right|, \text{ car} F_X(\mathbb{R}) \subseteq [0, 1]$$

avec égalité si et seulement si F_X est continue.
Au passage, on gagne le fait que la fonction de répartition limite $s \mapsto s$ est continue.
D'après la LGN, $\forall s \in [0, 1],\ \exists A_s \subseteq \Omega,\ \mathbb{P}(A_s) = 1$ telle que

$$\forall \omega \in A_s, \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{U_i(\omega) \leq s} - s \to 0$$

Par suite, $\exists A \subseteq \Omega,\ \mathbb{P}(A) = 1$ tel que $\forall s \in [0, 1] \cap \mathbb{Q}$ et

$$\forall \omega \in A,\ \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{U_i(\omega) \leq s} - s \to 0$$

Soit maintenant $s \in [0,1]$ quelconque. Il existe alors $s_k \to s$ en croissance et $t_k \to t$ en décroissante avec $s_k, t_k \in \mathbb{Q}$ et par monotonie

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{U_i(\omega)\leq s_k} \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{U_i(\omega)\leq s} \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{U_i(\omega)\leq t_k}$$

En faisant $n \to +\infty$, on déduit

$$s_k \leq \liminf_{n\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{U_i(\omega)\leq s} \leq \limsup_{n\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{U_i(\omega)\leq s} \leq t_k$$

On fait tendre $k \to +\infty$ et on conclut que

$$\forall \omega \in A,\ \forall s \in [0,1],\ \lim_{n\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{U_i(s)\leq s} - s = 0$$

presque sûrement

$$s \mapsto \frac{1}{n}\sum \mathbb{1}_{U_i \leq s} \text{ convergence vers } s \mapsto s$$

$\exists A \subseteq \Omega,\ \mathbb{P}(A) = 1$ tel que $\forall \omega \in A$

$$\left(s \mapsto \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{U_i(\omega)\leq s}\right) \to (s \mapsto s)$$

Par le corollaire, presque sûrement on a la convergence uniforme. □

2.8.10 Fonctions de répartition empirique : Théorème de Donsker

Ce théorème est la version TCL du théorème de Gliverko-Cantelli. On considère la suite de fonctions aléatoires

$$W_n(t) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{U_i \leq t} - t\right)$$

A t fixé, d'après le TCL on a

$$W_n(t) \to^{\mathcal{L}} \mathcal{N}(0, t(1-t))$$

On désigne par $D([0,1])$ l'espace des fonctions $[0,1] \to \mathbb{R}$ qui sont continue à droite limite à gauche.
On munit $D([0,1])$ de la norme uniforme $||.||_\infty$.

Théorème 2.8.7 (de Donsker)**.** *La suite de fonctions (W_n) converge en loi dans $(D([0,1]), ||.||_\infty)$ vers une fonction aléatoire limite continue W_∞, appelée "pont brownien".*
W_∞ est une fonction continue (mais dérivable nulle-part) aléatoire universelle telle que $W_\infty(0) = W_\infty(1) = 0$

Démonstration : Admis.

Théorème 2.8.8. *Lorsque $n \to +\infty$, on a*

$$\mathbb{P}(||W_n(.)||_\infty > \lambda) \to K(\lambda) = 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2\lambda^2 k^2} \textit{ (Loi de Kolmogorov)}, \ \forall \lambda > 0$$

On recueille des données $(x_i)_{1,...,n}$ et on aimerait savoir si ces données sont des réalisation de variables aléatoires indépendantes et identiquement distribuées.
Si $x_i = X_i(\omega)$ avec X_i indépendantes et identiquement distribuée

$$\left|\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i \leq t} - F_X(t)\right|\right|_\infty \to 0 \text{ presque sûrement}$$

si F_x continue

$$\sqrt{n}\left|\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i \leq t} - F_X(t)\right|\right|_\infty \to^{\mathcal{L}} ||W_\infty||_\infty$$

Si $((x_i) \neq (X_i(\omega)))$ avec X_i indépendantes et identiquement distribuée.

2.9 Vecteurs Gaussiens

On commence par quelques rappels

- On dit qu'une variable aléatoire réelle X est gaussienne $X \sim \mathcal{N}(m, \sigma^2)$ si elle admet la densité

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right), \text{ pour } x \in \mathbb{R}$$

- Si $X \sim \mathcal{N}(m, \sigma^2)$ alors

$$\varphi_X(t) = \mathbb{E}(e^{itX}) = e^{itm}e^{-\frac{\sigma^2 t^2}{2}}$$

- Si $X \sim \mathcal{N}(m, \sigma^2)$ alors $Y = \dfrac{X-m}{\sigma} \sim \mathcal{N}(0,1)$ et inversement si $Y \sim \mathcal{N}(0,1)$, $\sigma Y + m \sim \mathcal{N}(m, \sigma^2)$.
- Dans le cas où $\sigma^2 = 0$, on dit que $X \sim \mathcal{N}(m, \sigma^2)$ est dégénérée, alors $X = m$ presque sûrement.

2.9.1 Définitions et premières propriétés

Dans la suite, on notera les vecteurs $(x_1 \ \cdots \ x_d)^t$ et si $x, y \in \mathbb{R}^d$, $x.y = \sum_{i=1}^{d} x_i.y_i$.

Définition 2.9.1 (Vecteur aléatoire gaussien)**.** Un vecteur aléatoire $X = (X_1, ..., X_d)^t$ est gaussien si et seulement si pour tout $a \in \mathbb{R}^d$, $a.X = \sum_{i=1}^{d} a_i X_i$ est de loi gaussienne.

Remarque 2.2. Si $X = (X_1, ..., X_d)$ est gaussien alors

$$\forall i \in \{1, ..., d\},\ X_i \text{ est gaussien}$$

Pour s'en convaincre, il suffit de prendre $a = (0, ..., 0, 1, 0, ..., 0)$ avec le 1 en i-ième position.

Remarque 2.3. La réciproque est cependant fausse. En effet, on prend le contre-exemple suivant :

$$X \sim \mathcal{N}(0,1) \text{ et } \varepsilon \sim \mathcal{B}(\pm 1, \frac{1}{2})$$

Alors, en considérant le vecteur aléatoire $(X, \varepsilon X)$ dont les composantes suivent toutes les deux la gaussienne $\mathcal{N}(0,1)$, on se rend compte que ce n'est pas un vecteur gaussien.
En effet, si c'était le cas on devrait avoir $X + \varepsilon X$ suit une loi gaussienne.
Or, $X + \varepsilon X$ n'est pas constante

$$\mathbb{P}(X + \varepsilon X = 0) \geq \mathbb{P}(\varepsilon = -1) = \frac{1}{2}$$

C'est-à-dirre que $X + \varepsilon X$ a un atome ! On peut aussi énoncer un autre contre-exemple assez classique :

- $X \sim \mathcal{N}(0,1)$;
- $Y = X$ si $|X| \geq 1$, et $Y = -X$ sinon.
  Alors, $Y \sim \mathcal{N}(0,1)$.

Cependant (X, Y) n'est pas gaussien (considérer $X + Y$).

Remarque 2.4. Si X est un vecteur gaussien, $X \in L^2$. En effet par Cauchy-Schwarz on a

$$\mathbb{E}(|X_i X_j|) \leq \sqrt{\mathbb{E}(X_i^2)\mathbb{E}(X_j^2)} < \infty$$

Définition 2.9.2 (Moyenne d'un vecteur gaussien)**.** Si $X = (X_1, ..., X_d)^t$ est un vecteur gaussien, on définit sa moyenne par

$$m = (\mathbb{E}(X_1), ..., \mathbb{E}(X_d))^t$$

et sa matrice de covariance $K = (K_{i,j})_{1 \leq i,j \leq n}$ par

$$K_{i,j} = \mathrm{Cov}(X_i, X_j) = \mathbb{E}(X_i X_j - \mathbb{E}(X_i)\mathbb{E}(X_j))$$

Une première remarque sera de voir que la matrice de covariance K est symétrique et positive. En effet, si $X \in \mathbb{R}$, on a

$$x^t K x = \sum_{i,j=1}^{d} x_i x_j \mathrm{Cov}(X_i, X_j) = \mathrm{Var}\Big(\sum_{i=1}^{d} x_i X_i\Big) \geq 0$$

On va quand même détailler la dernière égalité qui a l'air de sortir de nul-part.

$$X \sim \mathcal{N}(0, K)$$

et,

$$\mathrm{Var}(x.X) = \mathrm{Var}\Big(\sum_{i=1}^{d} x_i X_i\Big),\ \sum_{i=1}^{d} x_i X_i \in \mathbb{R}$$

Par définition de X, on a que $\mathbb{E}(X) = 0$ et donc $\mathbb{E}(X_i) = 0$. Par linéarité de l'espérance on a donc juste à calculer

$$\mathbb{E}\Big(\Big(\sum_{i=1}^{d} x_i X_i\Big)^2\Big) = \mathbb{E}\Big(\sum_{i=1}^{d} x_i X_i \sum_{k=1}^{d} x_k X_k\Big) = \mathbb{E}\Big(\sum_{i=1}^{d}\sum_{k=1}^{d} x_i x_k X_i X_k\Big)$$

On utilise la linéarité de l'espérance

$$= \sum_{i=1}^{d}\sum_{k=1}^{d} x_i x_k \mathbb{E}(X_i X_k) = \sum_{i=1}^{d}\sum_{k=1}^{d} x_i x_k \mathrm{Cov}(X_i, X_k) = \sum_{i=1}^{d}\sum_{k=1}^{d} x_i x_k K_{i,k} = x^t K x$$

Proposition 2.9.1. *Un vecteur gaussien $X = (X_1, ..., X_d)^t$ de moyenne m et de covariance K a pour fonction caractéristique, pour $x = (x_1, ..., x_d)^t$*

$$\varphi_X(x) = \mathbb{E}(e^{ix.X}) = e^{ix.m} e^{-\frac{1}{2}x^t K x}$$

Démonstration : Par définition si X est gaussien et si $x \in \mathbb{R}^d$ alors $x.X$ est une variable aléatoire gaussienne avec

$$\mathrm{Var}(x.X) = \mathrm{Var}\Big(\sum_{i=1}^{d} x_i X_i\Big) = \sum_{i=1}^{d} x_i \mathbb{E}(X_i) = x.m$$

Par la remarque qui précède,

$$\mathrm{Var}(x.X) = \mathrm{Var}\Big(\sum_{i=1}^{d} x_i X_i\Big) = x^t K x$$

Ainsi la fonction caractéristique de $x.X$ prise en $t = 1$,

$$\mathbb{E}\Big(e^{itx.X}\Big) = e^{itx.m} e^{-\frac{t^2}{2}x^t K x}$$

□

Une deuxième remarque serait de voir que la loi d'un vecteur gaussien X est entièrement caractérisée par $m = \mathbb{E}(X)$ et la matrice de covariance K. On notera

$$X \sim \mathcal{N}_d(m, K)$$

Remarque 2.5. Les caractéristiques d'un vecteur gaussien se lisent sur la transformée de Fourier, par exemple

$$\varphi_X(s,t) = e^{2is+3it} e^{-\frac{1}{2}(s^2-2st+2t^2)}$$

$$m = (2,3)$$

$$K = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$$

Proposition 2.9.2. *Si $X \sim \mathcal{N}_d(m_x, K_X)$ et $Y \sim \mathcal{N}(m_Y, K_Y)$, et X et Y sont indépendants, alors $X+Y$ est gaussien*

$$X + Y \sim \mathcal{N}_d(m_X + m_Y, K_X + K_Y)$$

Démonstration : On calcule la fonction caractéristique

$$\varphi_{X+Y}(x)\mathbb{E}\left(e^{ix.(X+Y)}\right) = \mathbb{E}\left(e^{ix.X+ix.Y}\right) =^{\perp\!\!\!\perp} \mathbb{E}\left(e^{ix.X}\right)\mathbb{E}\left(e^{ix.Y}\right)$$

$$= \left(e^{ix.m_X}e^{-\frac{1}{2}x^tK_Xx}\right)\left(e^{ix.m_Y}e^{-\frac{1}{2}x^tK_Yx}\right) = e^{ix.(m_X+m_Y)}e^{-\frac{1}{2}x^t(K_X+K_Y)x}$$

□

Soit maintenant $(X, Y) = X_1, ..., X_d, Y_1, ..., Y_p) \in \mathbb{R}^{d+p}$ un vecteur gaussien, de moyenne $(m_X, m_Y)_i n\mathbb{R}^{d+p}$ et de matrice de covariance $K_{(X,Y)} \in \mathcal{M}_{d+p}(\mathbb{R})$.

Proposition 2.9.3. *Les vecteurs gaussiens X et Y sont indépendants si et seulement si $K_{(X,Y)}$ est diagonale par blocs, i.e.*

$$K_{(X,Y)} = \begin{pmatrix} K_X & 0 \\ 0 & K_Y \end{pmatrix}$$

avec $K_X \in \mathcal{M}_d(\mathbb{R})$ et $K_Y \in \mathcal{M}_p(\mathbb{R})$.

Démonstration : Le sens direct est clair.
Pour la réciproque il suffit de calculer la fonction caractéristique

$$\mathbb{E}\left(e^{i(x,y).(X,Y)}\right) = e^{i(x.m_X+y.m_Y)}e^{-\frac{1}{2}(x,y)^tK_{(X,Y)}(x,y)} = e^{i(x.m_X+y.m_Y)}e^{-\frac{1}{2}x^tK_Xx}e^{-\frac{1}{2}y^tK_Yy}$$

$$= \varphi_X(x)\varphi_Y(y)$$

□

Proposition 2.9.4. *Soit (X_n) une suite de vecteurs gaussiens dans $\mathbb{R}^d$, $X_n \sim \mathcal{N}_d(m_n, K_n)$. Alors si X_n converge en loi vers X, alors X est gaussien.*
Enfait,

$$X_n \to X \sim \mathcal{N}(m, K), \text{ où } m_n, K_n \to m, K$$

Démonstration : Plus tard. □

Proposition 2.9.5. *Soit $X \sim \mathcal{N}_d(m, K)$ et $A \in \mathcal{M}_{p,d}(\mathbb{R})$ alors $AX \sim \mathcal{N}_p(Am, AKA^T)$.*

Démonstration : Tout d'abord AX est gaussien car toute combinaison linéaire de ses composantes est une combinaison linéaire des X_i. Par définition, si $x \in \mathbb{R}^d$ alors

$$(AX).x = X.A^Tx$$

de sorte que

$$\mathbb{E}((AX).x) = \mathbb{E}(X).A^Tx = m.A^Tx = Am.x$$

$$\mathrm{Var}((AX).x) = \mathrm{Var}(X.A^Tx) = (A^Tx)^TK(A^Tx)x^T(AKA^T)x$$

Si bien que $\mathbb{E}(AX) = Am$ et $\mathrm{Cov}(AX) = AKA^T$. □

Définition 2.9.3 (Non-degénérescence)**.** On dit qu'un vecteur gaussien $X \sim \mathcal{N}_d(m, K)$ est non-degénéré si sa matrice de covariance est inversible. C'est-à-dire, si $\det(K) \neq 0$.
Dans le cas inverse, on dira que X est dégénéré.

Faisons maintenant quelques observations.

On a dit que $X \sim \mathcal{N}_d(m, K)$ est dégénéré si $\det(K) = 0$. Donc enfait, s'il existe $a \in \mathbb{R}^d$ tel que $K.a = 0$. Alors $\mathrm{Var}(AX) = a^T K a = 0$, de sorte que AX soit constant presque sûrement.
Par exemple, si on prend $Y_1, ..., Y_{d-1} \sim \mathcal{N}(0, 1)$ indépendants et on construit le vecteur gaussien suivant :

$$X = \left(Y_1, ..., Y_{d-1}, -\sum_{i=1}^{d-1} Y_i \right)$$

La dernière colonne étant une combinaison linéaire des $d-1$ précédente, le déterminant de la comatrice de ce vecteur gaussien est nul. D'où la dégénérescence.

Maintenant, prenons le cs où $X \sim \mathcal{N}_d(m, K)$ non-dégénéré. Alors K est symétrique définie positive. Par le théorème spectrale elle est diagonalisable en base orthonormée, ie

$$\exists P \text{ orthonormée} \mid K = PDP^T$$

avec

$$D = \mathrm{diag}(\lambda_1, ..., \lambda_d) \text{ avec } \lambda_i > 0$$

Ainsi, K admet une racine carré matricielle

$$\sqrt{K} = P\mathrm{diag}(\sqrt{\lambda_1}, ..., \sqrt{\lambda_d})P^T$$

$$(\sqrt{K})^{-1} = P\mathrm{diag}\left(\frac{1}{\sqrt{\lambda_1}}, ..., \frac{1}{\sqrt{\lambda_d}} \right) P^T$$

Proposition 2.9.6. *Si $X \sim \mathcal{N}_d(m, K)$ est un vecteur gaussien non-dégénéré, alors*

$$(\sqrt{K})^{-1}(X - m) \sim \mathcal{N}_d(0, I_d)$$

Démonstration : D'après ce que l'on a fait précédemment,

$$\mathbb{E}((\sqrt{K})^{-1}(X - m)) = 0$$

$$\mathrm{Var}((\sqrt{K})^{-1}(X - m)) = (\sqrt{K})^{-1} K ((\sqrt{K})^{-1})^T = I_d$$

□

Proposition 2.9.7. *Soit $X \sim \mathcal{N}_d(m, K)$ un vecteur gaussien non-dégénéré. Alors, X admet la densité suivante par rapport à la mesure de Lebesgue dans $\mathbb{R}^d$*

$$f_X(x_1, ..., x_d) = \frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{\det(K)}} e^{-\frac{1}{2}(x-m)^T K^{-1}(x-m)}$$

Démonstration : Si $K = I_d$, alors $X = (X_1, ..., X_d)$ avec les X_i indépendants et indentiquement distribués $\mathcal{N}(0, 1)$. Dans ce cas la densité f_X de X est simplement le produit des densité d'une $\mathcal{N}(0, 1)$. Pour passer au cas K quelconque, on effectue le changement de variable $x = (\sqrt{K})^{-1}(x - m)$. □

3 Conditionnement (M1)

Sources : **[1]** Chaînes de Markov, Jean-Christophe Breton, Université de Rennes - ENS Rennes, 2023-2024.
[2] Probabilités, Ying Hu, Université de Rennes - ENS Rennes, 2023-2024.

On introduit la notion de conditionnement dans un cadre élémentaire discret pour commencer. L'approche plus générale sera l'objet de la deuxième partie. On considère un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$.

3.1 Conditionnement discret

3.1.1 Probabilité conditionnelle discrète

Définition 3.1.1 (Probabilité conditionnelle)**.** Soit B un évènement de probabilité non-nulle. Pour tout évènement A, on définit

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Si $A \perp\!\!\!\perp B$, $\mathbb{P}(A|B) = \mathbb{P}(A)$.

Proposition 3.1.1. *Soit $B \in \mathcal{F}$ avec $\mathbb{P}(B) > 0$. La fonction d'ensemble*

$$\mathbb{P}(*|B) : A \in \mathcal{F} \mapsto \mathbb{P}(A|B)$$

est une probabilité sur $(\Omega, \mathcal{F})$.

Démonstration : Il est clair que $\mathbb{P}(A|B) \geq 0$ et que $\mathbb{P}(\emptyset|B) = 0$. Puis la σ-additivité découle de celle de $\mathbb{P}$:

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i \Big| B\Big) = \frac{1}{\mathbb{P}(B)}\mathbb{P}\Big(\Big(\bigcup_{i=1}^{\infty} A_i\Big) \cap B\Big) = \frac{1}{\mathbb{P}(B)}\mathbb{P}\Big(\bigcup_{i=1}^{\infty}(A_i \cap B)\Big)$$

$$= \sum_{i=1}^{\infty} \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \sum_{i=1}^{\infty} \mathbb{P}(A_i|B)$$

Finalement,

$$\mathbb{P}(\Omega|B) = \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1$$

□

Cette propriété assurre que l'on dispose, pour les probabilités conditionnelles, de toutes les propriétés habituelles sur les probabilités. On dispose aussi de quelques propriétés spécifiques.

Proposition 3.1.2 (Règle des conditionnements successifs)**.** *Soient n évènements $A_1, ..., A_n$ tels que $\mathbb{P}(A_1 \cap ... \cap A_{n-1}) \neq 0$. Alors*

$$\mathbb{P}\Big(\bigcap_{i=1}^{n}\Big) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)...\mathbb{P}(A_n|A_1 \cap ... \cap A_{n-1})$$

Démonstration : Il suffit de développer le membre de droite avec la formule des probabilités conditionnelles et de simplifier. □

Proposition 3.1.3. *Soient A, B, C des évènements tels que $\mathbb{P}(B \cap C) > 0$. En notant $\mathbb{P}_C = \mathbb{P}(*|C)$ alors*

$$\mathbb{P}_C(A|B) = \mathbb{P}(A|B \cap C)$$

Démonstration : De même on utilise la formule des probabilités conditionnelles

$$\mathbb{P}_C(A|B) = \frac{\mathbb{P}_C(A \cap B)}{\mathbb{P}_C(B)} = \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(C)} \frac{\mathbb{P}(C)}{\mathbb{P}(B \cap C)} = \mathbb{P}(A|B \cap C)$$

□

Enchainer les conditionnements revient donc à conditionner par l'intersection.

Dans l suite on note $I \subseteq \mathbb{N}$ pour désigner un ensemble dénombrable (fini ou infini).

Définition 3.1.2 (Système complet)**.** On appelle système complet d'évènements toute suite dénombrable $(B_i)_{i \in I}$ d'évènements deux à deux disjoints et tels que

$$\sum_{i \in I} \mathbb{P}(B_i) = 1$$

Théorème 3.1.1 (Formule des probabilités totales)**.** *Soit (B_i) un système complet dénombrable de Ω avec $\mathbb{P}(B_i) > 0$ pour tout i.Pour tout $A \in \mathcal{F}$, on a*

$$\mathbb{P}(A) = \sum_{i \in I} \mathbb{P}(A|B_i)\mathbb{P}(B_i)$$

Démonstration : Notons $\Omega_0 = \sqcup_{i \in I} B_i$. Alors comme les (B_i) sont un SC, $\mathbb{P}(\Omega_0) = 1$. Comme les $(A \cap B_i)$ sont disjoints (car les (B_i) le sont),

$$\mathbb{P}(A) = \mathbb{P}(A \cap \Omega_0) = \mathbb{P}\left(\bigcap_{i \in I}(A \cap B_i)\right) =_s um_{i \in i}\mathbb{P}(A \cap B_i) = \sum_{i \in I} \mathbb{P}(A|B_i)\mathbb{P}(B_i)$$

□

Quand on connait les $\mathbb{P}(A|B_i)$ pour tout un système de partition (B_i), on peut chercher les conditionnements inverses $\mathbb{P}(B_i|A)$.

Théorème 3.1.2 (Formule de Bayes)**.** *Soit (B_i) un SC de Ω avec $\mathbb{P}(B_i) > 0$ pour tout i. Pour tout évènement A tel que $\mathbb{P}(A) > 0$,*

$$\forall j \in I,\ \mathbb{P}(B_j|A) = \frac{\mathbb{P}(A|B_j)\mathbb{P}(B_j)}{\sum_{i \in I} \mathbb{P}(A|B_i)\mathbb{P}(B_i)}$$

Démonstration : On applique les probabilités conditionnelles à $\mathbb{P}(B_j|A)$ puis la formule des probabilités conditionnelles au dénominateur. □

Cette formule est à l'origine de tout un pan des statistiques (dites bayésiennes) qui consiste à inverser des conditionnements en manipulant des probabilités dites a priori ou a posteriori.

Maintenant, si on prend $B = \{Y = y\}$ avec Y une variable aléatoire discrète et y un atome, on donne un sens à

$$\mathbb{P}(A|Y = y) = \frac{\mathbb{P}(A, Y = y)}{\mathbb{P}(Y = y)}$$

Maintenant on peut aussi conditionner selon une variable aléatoire Y et ainsi $\mathbb{P}(A|Y)$ devient aussi une variable aléatoire.

Définition 3.1.3 (Probabilité conditionnelle discrète)**.** Etant donné une variable aléatoire discrète Y de support $\mathcal{S}(Y) = \{y_j | j \in J\}$, on appelle probabilité conditionnelle selon Y la fonction d'ensemble

$$\begin{array}{rccc} \mathbb{P}(*|Y) & : & \mathcal{F} & \to & [0,1]^\Omega \\ & & A & \mapsto & \sum_{j\in J} \mathbb{P}(A|Y = y_j)\mathbb{1}_{\{Y=y_j\}} \end{array}$$

Si bien que, sur l'évènement $\{Y = y\}$, $\mathbb{P}(A|Y) = \mathbb{P}(A|Y = y)$.

Dans le cas plus général où Y est à densité, c'est plus compliqué car les conditionnements par $\{Y = y\}$ sont singuliers (évènements négligeables).

3.1.2 Espérance conditionnelle discrète

Etant donné B non négligeable, on définit l'espérance conditionnelle sachant B comme l'espérance par rapport à la probabilité $\mathbb{P}(*|B)$.

Définition 3.1.4. Soit X une variable positive ou L^1,

$$\mathbb{E}(X|B) = \int_\Omega X(\omega)\mathbb{P}(d\omega|B)$$

Proposition 3.1.4. *Soit X une variable intégrable et B non-négligeable.*

$$\mathbb{E}(X|B) = \frac{\mathbb{E}(X\mathbb{1}_B)}{\mathbb{P}(B)}$$

Démonstration : On suit la méthode habituelle. Pour $X = \mathbb{1}_A$, il s'agit de la définition de $\mathbb{P}(A|B)$ car $\mathbb{E}(X\mathbb{1}_B) = \mathbb{E}(\mathbb{1}_A\mathbb{1}_B) = \mathbb{E}(\mathbb{1}_{A\cap B}) = \mathbb{P}(A\cap B)$. Puis, par linéarité de l'espérance, la formule est vraie pour les fonctions étagées positives

$$X = \sum_{i=1}^{n} \alpha_i \mathbb{1}_{A_i},\ \alpha_i \geq 0$$

Par convergence monotone, la formule est vraie pour $X \geq 0$.
Si bien qu'en notant $X = X^+ - X^-$, la formule est vraie pour X de signe quelconque.
La différence $\mathbb{E}(X^+\mathbb{1}_B) - \mathbb{E}(X\mathbb{1}_B)$ a bien un sens car X est intégrable.□

Si X est une variable aléatoire discrète de support $\mathcal{S}(X) = \{x_i | i \in I\}$, alors

$$X = \sum_{i \in I} x_i \mathbb{1}_{\{X = x_i\}}$$

On a

$$\mathbb{E}(X|B) = \sum_{i \in I} x_i \mathbb{P}(X = x_i|B)$$

Dans le cadre où Y est une autre variable aléatoire discrète de support $\mathcal{S}(Y)$, on définit de cette façon

$$\mathbb{E}(X|Y = y_j) = \sum_{i \in I} x_i \mathbb{P}(X = x_i|Y = y_j)$$

et comme fait précédemment, on peut généraliser

Définition 3.1.5 (Espérance conditionnelle discrète)**.** Soit X une variable aléatoire intégrable et Y une variable aléatoire discrète. L'espérance conditionnelle de X sachant Y est définie par

$$\mathbb{E}(X|Y) = \sum_{j \in J} \mathbb{E}(X|Y = y_j) \mathbb{1}_{\{Y = y_j\}}$$

i.e. $\mathbb{E}(X|Y) = \mathbb{E}(X|Y = y_j)$ sur l'évènement $\{Y = y_j\}$.

Il est important de comprendre que $\mathbb{E}(X|Y = y_j) \in \mathbb{R}$ alors que $\mathbb{E}(X|Y)$ est une variable aléatoire.

En combinant les formules précédentes, on a

$$\mathbb{E}(X|Y) = \sum_{(i,j) \in I \times J} x_i \mathbb{P}(X = x_i|Y = y_j) \mathbb{1}_{\{Y = y_j\}}$$

On sait que $\mathbb{E}(\mathbb{1}_A|B) = \mathbb{P}(A|B)$. Cependant, on a aussi le lien naturel

$$\mathbb{E}(\mathbb{1}_A|Y) = \mathbb{P}(A|Y)$$

On vérifie aussi rapidement que $\mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}(X)$. En effet, on a que $\mathbb{E}(X|B)\mathbb{P}(B) = \mathbb{E}(X\mathbb{1}_B)$ et donc par linéarité de l'espérance

$$\mathbb{E}(\mathbb{E}(X|Y)) = \sum_{j \in J} \mathbb{E}(X|Y = y_j)\mathbb{P}(Y = y_j) = \sum_{j \in J} \mathbb{E}(X\mathbb{1}_{\{Y = y_j\}}) = \mathbb{E}(X)$$

Plus généralement, on la propriété suivante :

Proposition 3.1.5 (Conditionnement en cascade)**.** *Soient X, Y, Z des variables aléatoires discrètes. On a*

$$\mathbb{E}(X|Y) = \mathbb{E}(\mathbb{E}(X|Y, Z)|Y)$$

Démonstration : Pus tard. □

4 Chaînes de Markov (M1)

Sources : **[1]** Chaînes de Markov, Jean-Christophe Breton, Université de Rennes - ENS Rennes, 2023-2024.
[2]

4.1 Dynamique markovienne

4.1.1 Probabilité de transition

4.1.2 Exemples de chaînes de Markov

4.1.3 Probabilités trajectorielles

4.1.4 Chaîne de Markov canonique

4.1.5 Propriétés de Markov

4.2 Récurrence et transience

4.2.1 Etats récurrents et transitoires

4.2.2 Ensembles clos et irréductibilité

4.2.3 Classes de récurrence

4.2.4 Absorption dans les classes de récurrences

4.3 Invariance et équilibre

4.3.1 Mesures invariantes

4.3.2 Invariance et récurrence

4.3.3 Périodicité et forte irréductibilité

4.3.4 Equilibre d'une chaine de Markov

4.3.5 Théorème ergodique

5 Martingales (M1)

Sources : **[1]** Martingales, Jean-Christophe Breton, Université de Rennes - ENS Rennes, 2023-2024.
[2] Probabilités et Martingales, Ying Hu, Université de Rennes - ENS Rennes, 2023-2024.

On va introduire la notion de Martingales, concept fondamental en théorie des probabilités et en statistiques. Elles sont utilisées pour modéliser des systèmes où les valeurs futures ne peuvent pas être prédites à partir de valeurs passées, par exemple dans les marchés financiers où les prix des actifs sont souvent considérés comme des martingales sous certaines hypothèses. On verra qu'elles jouent un rôle crucial dans les théorèmes de convergence des variables aléatoires. On peut aussi noter que les martingales sont un type particulier de processus stochastique, elles sont souvent utilisées pour étudier des processus plus complexes comme les chaînes de Markov et les processus de diffusion. Elles sont aussi étroitement liées à l'intégrale stochastique, par exemple l'intégrale d'Itô.

Dans la suite, on considère un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$.

5.1 Martingales et filtrations

5.1.1 Filtration et mesurabilité

Définition 5.1.1 (Filtration)**.** Soit $(\mathcal{F}_n)_{n\geq 0}$ une suite de sous-tribus de $\mathcal{F}$. On dit que $(\mathcal{F}_n)_n$ est une filtration lorsque pour tout $n \geq 0$, on a $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$.
Un espace probabilisé muni d'une filtration est appelé espace de probabilité filtrée.

En terme d'informations, une filtration peut être interprétée comme une quantité d'information qui évolue au cours du temps : $\mathbb{N}$ est le temps et $\mathcal{F}_n$ est l'information disponible à la date n.

Définition 5.1.2 (Adapté)**.** On dit qu'une suite $(X_n)_n$ est adaptée par rapport à une filtration $(\mathcal{F}_n)$ si pour tout $n \geq 0$, X_n est $\mathcal{F}_n$-mesurable.

Ainsi, une suite (X_n) est adaptée si X_n est connue à la date n.

Exemple 5.1 (Filtration canonique). Si $(X_n)_{n\geq 1}$ est une suite de variables aléatoires, on appelle filtration canonique ou naturelle la filtration $\mathcal{F}_n$ des tribus engendrées par ces variables aléatoires :

$$\mathcal{F}_n = \sigma(X_1, ..., X_n) = \sigma\left(\bigcup_{i=1}^{n} \sigma(X_i)\right),\ n \geq 1$$

Par construction, la suite $(X_n)_{n\geq 1}$ est adaptée par rapport à sa filtration naturelle.

Exemple 5.2 (Filtration dyadique). Soit $(\Omega, \mathcal{F}, \mathbb{P}) = (]0,1[, \mathcal{B}(]0,1[), \lambda)$. On pose

$$\mathcal{F}_n = \sigma\left(\left[\frac{i-1}{2^n}, \frac{i}{2^n}\right[, i = 1, ..., 2^n\right)$$

On a bien une filtration car pour tout $n \geq 1$ et $1 \leq i \leq 2^n$

$$\left[\frac{i-1}{2^n}, \frac{i}{2^n}\right[= \left[\frac{2i-2}{2^{n+1}}, \frac{2i-1}{2^{n+1}}\right[\cup \left[\frac{2i-1}{2^{n+1}}, \frac{2i}{2^{n+1}}\right[$$

On a

$$\mathcal{D}_n = \left\{ \left[\frac{i-1}{2^n}, \frac{i}{2^n} \right[, i = 1, ..., 2^n \right\} \subseteq \sigma(\mathcal{D}_{n+1})$$

Et donc

$$\mathcal{F}_n = \sigma(\mathcal{D}_n) \subseteq \sigma(\mathcal{D}_{n+1}) = \mathcal{F}_{n+1}$$

Définition 5.1.3 (Prévisibilité)**.** Une suite de variables aléatoires (H_n) est dite prévisible pour une filtration $(\mathcal{F}_n)$ si pour tout $n \geq 1$, H_n est $\mathcal{F}_{n-1}$-mesurable.

Ainsi, une suite (H_n) est prévisible si H_n peut être prédite avec l'information $\mathcal{F}_{n-1}$ disponible à la date $n-1$.

5.1.2 Temps d'arrêt

Définition 5.1.4 (Temps d'arrêt)**.** Une variable aléatoire T à valeurs dans $\mathbb{N} \cup \{+\infty\}$ est un $(\mathcal{F}_n)$-temps d'arrêt si pour tout $n \geq 0$, on a $\{T \leq n\} \in \mathcal{F}_n$.

Exemple 5.3. Si T est constant égal à n_0, alors T est un temps d'arrêt.

Exemple 5.4 (Temps d'atteinte). Soit (X_n) une suite de variables aléatoires et $(\mathcal{F}_n)$ sa filtration naturelle.

$$T = \min\{i \geq 0 | X_i \in A\}$$

est un $(\mathcal{F}_n)$-temps d'arrêt pour $A \in \mathcal{B}(\mathbb{R})$. En effet,

$$\{T \leq n\} = \bigcup_{k=0}^{n} \{X_k \in A\} \in \mathcal{F}_n$$

car $\{X_k \in A\} \in \mathcal{F}_k \subseteq \mathcal{F}_n$.

Remarque 5.1. Etant donné un $(\mathcal{F}_n)$-temps d'arrêt T, on définit une suite prévisible par

$$H_n = \mathbb{1}_{\{T \geq n\}}$$

Proposition 5.1.1 (Propriétés des temps d'arrêt)**.** *Quelques propriétés importantes*

1. *T est un temps d'arrêt si et seulement si pour tout $n \geq 0$ on a $\{T = n\} \in \mathcal{F}_n$;*
2. *Si T et S sont des $(\mathcal{F}_n)$-temps d'arrêt. Alors $T \wedge S = \min(T, S)$, $T \vee S = \max(T, S)$, $T + S$ en sont aussi ;*
3. *Si T est un $(\mathcal{F}_n)$-temps d'arrêt alors pour tout $k \geq 0$, $T \wedge k$ en est un aussi ;*
4. *Si (T_p) est une suite monotone de $(\mathcal{F}_n)$-temps d'arrêt alors $T = \lim T_p$ est aussi un temps d'arrêt ;*
5. *Soit (T_p) est une suite de $(\mathcal{F}_n)$-temps d'arrêt alors*

$$\inf T_p, \ \sup T_p, \ \liminf T_p, \ \limsup T_p$$

sont des $(\mathcal{F}_n)$-temps d'arrêt.

Démonstration : **1.** Si T est un $(\mathcal{F}_n)$-temps d'arrêt, alors

$$\{T = n\} = \{T \leq n\} \backslash \{T \leq n-1\} \in \mathcal{F}_n$$

car $\{T \leq n\} \in \mathcal{F}_n$ et $\{T \leq n-1\} \in \mathcal{F}_{n-1} \subseteq \mathcal{F}_n$.
La réciproque vient du fait que

$$\{T \leq n\} = \bigcup_{k \leq n} \{T = k\} \in \mathcal{F}_n$$

car $\{T = k\} \in \mathcal{F}_k \subseteq \mathcal{F}_n$.
2. En effet, pour $n \in \mathbb{N}$,

$$\{T \wedge S \leq n\} = \{T \leq n\} \cup \{S \leq n\} \in \mathcal{F}_n$$

$$\{T \vee S \leq n\} = \{T \leq n\} \cap \{S \leq n\} \in \mathcal{F}_n$$

$$\{T + s = n\} = \bigcup_{k=0}^{n} \{T = k\} \cap \{S = n-k\} \in \mathcal{F}_n$$

3. Cela découle du point précédent avec $S = k$ qui est un temps d'arrêt.
On peut aussi le faire directement, pour $n \geq k$, $\{T \wedge k \leq n\} = \Omega \in \mathcal{F}_n$ et pour $n < k$, on a

$$\begin{aligned}\{T \wedge k \leq n\} &= (\{T \wedge k \leq n\} \cap \{T \leq k\}) \cup (\{T \wedge k \leq n\} \cap \{T > k\}) \\ &= (\{T \leq n\} \cap \{T \leq k\}) \cup (\{k \leq n\} \cap \{T > k\}) \\ &= \{T \leq k \wedge n\} \cup (\{T \leq k\}^c \cap \emptyset) = \{T \leq k \wedge n\} \in \mathcal{F}_{k \wedge n} \subseteq \mathcal{F}_n\end{aligned}$$

4. Pour tout $n \in \mathbb{N}$, on a dans les cas croissant et décroissant respectivement

$$\{T \leq n\} = \left\{ \lim_{p \to \infty} T_p \leq n \right\} = \bigcap_{p \leq 1} \{T_p \leq n\} \in \mathcal{F}_n$$

$$\{T \leq n\} = \left\{ \lim_{p \to \infty} T_p \leq n \right\} = \bigcup_{p \leq 1} \{T_p \leq n\} \in \mathcal{F}_n$$

5. Cela découle des propriétés précédentes en écrivant

$$\inf T_p = \lim_n \min_{1 \leq p \leq n} T_p, \ \liminf T_p = \sup_n \inf_{k \geq n} T_k$$

$$\sup T_p = \lim_n \max_{1 \leq p \leq n} T_p, \ \limsup T_p = \inf_n \sup k \geq n T_k$$

ou alors on peut directement le déduire via

$$\left\{ \inf_p T_p \leq n \right\} = \bigcup_{p \geq 1} \{T_p \leq n\}, \ \left\{ \max_p T_p \leq n \right\} = \bigcap_{p \geq 1} \{T_p \leq n\}$$

$$\left\{ \liminf_p T_p \leq n \right\} = \bigcup_{m \geq 0} \bigcap_{p \geq m} \{T_p \leq n\}, \left\{ \limsup_p T_p \leq n \right\} = \bigcap_{m \geq 0} \bigcup_{p \geq m} \{T_p \leq n\}$$

□

Exemple 5.5. On va prendre des exemples du TD.

Définition 5.1.5 (Tribu d'un temps d'arrêt)**.** A un temps d'arrêt T, on associe la tribu

$$\mathcal{F}_T = \{A \in \mathcal{F} | \forall n \in \mathbb{N}, A \cap \{T \leq n\} \in \mathcal{F}_n\}$$

Proposition 5.1.2. *Lorsque T est un temps d'arrêt, $\mathcal{F}_T$ est bien une tribu.*

Démonstration : Tout d'abord, $\Omega \cap \{T \leq n\} = \{T \leq n\} \in \mathcal{F}_n$ pour tout n et donc $\Omega \in \mathcal{F}_T$.
Puis, si $A_i,\ i \in I \subseteq \mathbb{N}$ sont dans $\mathcal{F}_T$ alors

$$\left(\bigcup_{i\in I} A_i\right) \cap \{T \leq n\} = \bigcup_{i\in I}(A_i \cap \{T \leq n\}) \in \mathcal{F}_n$$

car $A_i \cap \{T \leq n\} \in \mathcal{F}_n$ (car $A_i \in \mathcal{F}_T$).
Finalement, si $A \in \mathcal{F}_T$ alors pour tout $n \geq 0$,

$$A^c \cap \{T \leq n\} = \{T \leq n\} \backslash (A \cap \{T \leq n\}) \in \mathcal{F}_n$$

car $A \cap \{T \leq n\} \in \mathcal{F}_n$ et $\{T \leq n\} \in \mathcal{F}_n$. Cela assure $A^c \in \mathcal{F}_T$. □

On remarque que $\sigma(T) \subseteq \mathcal{F}_T$ donc T est en particulier $\mathcal{F}_T$-mesurable.

De façon générale, en suivant les remarques précédentes, on peut interpréter $\mathcal{F}_T$ comme l'information disponible à la date aléatoire T.

Proposition 5.1.3 (Propriétés des tribus $\mathcal{F}_T$)**.** *Voici les propriétés essentielles de la tribu $\mathcal{F}_T$*

1. *Pour un temps d'arrêt constant $T = n_0$ alors on a bien $\mathcal{F}_T = \mathcal{F}_{n_0}$;*
2. *Si $T \leq S$ sont deux temps d'arrêts alors $\mathcal{F}_T \subseteq \mathcal{F}_S$;*
3. *Un temps d'arrêt T est $\mathcal{F}_T$-mesurable ;*
4. *Pour T, S deux temps d'arrêts, on a $\mathcal{F}_{T\wedge S} = \mathcal{F}_T \cap \mathcal{F}_S$. De plus, $\{T \leq S\},\ \{S \leq T\},\ \{T = S\} \in \mathcal{F}_{T\wedge S}$;*
5. *Pour $A \in \mathcal{F}$ et T un temps d'arrêt, posons $T_A(\omega) = T(\omega)$ si $\omega \in A$, $T_A(\omega) = +\infty$ sinon. Alors $a \in \mathcal{F}_T$ si et seulement si T_A est un temps d'arrêt.*

Démonstration : Plus tard. □

Proposition 5.1.4. *Soit (X_n) une suite $(\mathcal{F}_n)$-adaptée et T un temps d'arrêt. Alors la variable aléatoire*

$$\mathbb{1}_{\{T<+\infty\}} X_T = \begin{cases} X_n, & \text{si } T = n \\ 0, & \text{si } T = +\infty \end{cases}$$

est $\mathcal{F}_T$-mesurable. Lorsque $T < +\infty$ p.s., il n'y a pas d'ambiguité de notations et on écrit simplement X_T.

Démonstration : Plus tard. □

5.1.3 Martingales, sous-martingales et sur-martingales

Définition 5.1.6 (Martingale)**.** Une suite de variables aléatoires (X_n) est une martingale par rapport à une filtration $(\mathcal{F}_n)$ si

1. $\mathbb{E}(|X_n|) < +\infty$ pour tout $n \geq 0$;
2. La suite (X_n) est $(\mathcal{F}_n)$-adaptée ;
3. Pour tout $n \geq 0$,
$$\mathbb{E}(X_{n+1}|\mathcal{F}_n) = X_n$$

Définition 5.1.7 (Sous-martingale et Sur-martingale)**.** Il n'y a que le point 3. qui change :

- Sous-martingale : $\mathbb{E}(X_{n+1}|\mathcal{F}_n) \geq X_n$ pour tout $n \geq 0$;
- Sur-martingale : $\mathbb{E}(X_{n+1}|\mathcal{F}_n) \leq X_n$ pour tout $n \geq 0$.

Exemple 5.6 (Marche aléatoire). Soit (X_n) une suite de variables aléatoires intégrables indépendantes centrées alors
$$S_n = \sum_{i=1}^{n} X_i,\ n \geq 1,\ S_0 = 0$$
est une (sur/sous)-martingale par rapport à la filtration naturelle $\mathcal{F}_n = \sigma(X_1, ..., X_n)$, avec $\mathcal{F}_0 = \{\emptyset, \Omega\}$, selon le signe ou la nullité de $\mathbb{E}(X_1)$.
En effet, S_n est clairement $\mathcal{F}_n$-mesurable et intégrable car les X_i le sont. Comme $X_{n+1} \perp\!\!\!\perp \mathcal{F}_n$, il vient
$$\mathbb{E}(S_{n+1}|\mathcal{F}_n) = \mathbb{E}(S_n + X_{n+1}|\mathcal{F}_n) = S_n + \mathbb{E}(X_{n+1}|\mathcal{F}_n) = S_n + \mathbb{E}(X_{n+1}) = S_n$$
dans le cas centré (on adapte facilement aux cas $\mathbb{E}(X_1) > 0$ et $\mathbb{E}(X_1) < 0$).

Remarque 5.2. Dans l'exemple précédent,
$$X_n = S_n - S_{n-1} = S_n - \mathbb{E}(S_n|\mathcal{F}_{n-1})$$
Ainsi, la suite (X_n) n'est pas une martingale mais une différence de martingale.
De manière générale, une suite de variables aléatoires indépendantes est une différence de martingale.

Exemple 5.7 (Modèle auto-régressif). Soit (ε_n) une suite de v.a. i.i.d. intégrables centrées et $a > 0$. On pose
$$X_{n+1} = aX_n + \varepsilon_{n+1},\ n \geq 0,\ X_0 = x$$
Alors, la variable aléatoire
$$Y_n = \frac{X_n}{a^n},\ n \geq 0$$
forme une martingale par rapport à a filtration naturelle $\mathcal{F}_n = \sigma(\varepsilon_1, ..., \varepsilon_n)$, $n \geq 1$, $\mathcal{F}_0 = \{\emptyset, \Omega\}$.
En effet, par récurrence, les Y_n sont intégrables et $\mathcal{F}_n$-mesurables, $n \geq 1$. Puis
$$\mathbb{E}(Y_{n+1}|\mathcal{F}_n) = \frac{1}{a^{n+1}}\mathbb{E}(aX_n + \varepsilon_{n+1}|\mathcal{F}_n) = \frac{1}{a^{n+1}}(aX_n + \mathbb{E}(\varepsilon_{n+1})) = \frac{1}{a^n}X_n = Y_n$$
car X_n est $\mathcal{F}_n$-mesurable et $\varepsilon_{n+1} \perp\!\!\!\perp \mathcal{F}_n$.

Maintenant, voyons un exemple très important dans le cadre de ce cours car il se reporte au cours de Modèles Aléatoires.

Exemple 5.8 (Galton-Watson). Soit $(X_{i,j})$ une famille de variables aléatoires entières i.i.d. de loi μ (sur $\mathbb{N}$) admettant pour moyenne m. On pose $Z_0 = 1$ et pour $n \geq 1$

$$Z_{n+1} = \sum_{j=1}^{Z_n} X_{n+1,j}$$

Alors, $(Z_n \backslash m^n)_n$ est une martingale par rapport à la filtration donnée par $\mathcal{F}_n = \sigma(X_{i,j} | i \leq n, j \geq 1)$. D'abord, on observe par récurrence que Z_n est $\mathcal{F}_n$-mesurable. En effet, si Z_n l'est alors pour tout $A \in \mathbb{N}$,

$$\{Z_{n+1} \in A\} = \bigcup_{p=0}^{+\infty} \{Z_{n+1} \in A, Z_n = p\} = \bigcup_{p=0}^{+\infty} \left(\left\{ \sum_{j=1}^{p} X_{n+1,j} \in A \right\} \cap \{Z_n = p\} \right) \in \mathcal{F}_{n+1}$$

Finalement, la propriété de martingale est satisfaite

$$\mathbb{E}(Z_{n+1} | \mathcal{F}_n) = \mathbb{E}\left(\sum_{j=1}^{Z_n} X_{n+1,j} \middle| \mathcal{F}_n \right) = \sum_{j=1}^{Z_n} \mathbb{E}(X_{n+1,j} | \mathcal{F}_n) = Z_n \mathbb{E}(X_{n+1,j}) = Z_n m$$

car Z_n est $\mathcal{F}_n$-mesurable et $X_{n+1} \perp\!\!\!\perp \mathcal{F}_n$. Si bien que

$$\mathbb{E}\left(\frac{Z_{n+1}}{m^{n+1}} \middle| \mathcal{F}_n \right)$$

Cette martingale modélise l'évolution d'une population avec loi de reproduction μ. Pour plus d'informations, se reporter aux cours de Modèles Aléatoires.

5.1.4 Propriétés des martingales

De façon générale, les énoncés pour les martingales s'adaptent pour des sous-martingales ou des sur-martingales.

Proposition 5.1.5. *Si (X_n) est une $\mathcal{F}_n$-martingale alors (X_n) est une $(\mathcal{G}_n)$-martingale pour $\mathcal{G}_n = \sigma(X_1, ..., X_n)$.*

Démonstration : Puisque (X_n) est une $(\mathcal{F}_n)$-martingale, chaque X_n est intégrable (car $\mathbb{E}(|X_n|) < \infty$). Puis par définition, (X_n) est $(\mathcal{G}_n)$-adaptée. Comme $X_1, ..., X_n$ sont $\mathcal{F}_n$-mesurables, il est immédiat que $\mathcal{G}_n \subseteq \mathcal{F}_n$.
Par le théorème de conditionnement en cascade,

$$\mathbb{E}(X_{n+1} | \mathcal{G}_n) = \mathbb{E}(\mathbb{E}(X_{n+1} | \mathcal{F}_n) | \mathcal{G}_n) = \mathbb{E}(X_n | \mathcal{G}_n) = X_n$$

Puisque, (X_n) est une $\mathcal{F}_n$-martingale donc $\mathbb{E}(X_{n+1} | \mathcal{F}_n) = X_n$ et X_n est $\mathcal{G}_n$-mesurable.□

Proposition 5.1.6. *Dans la définition d'une martingale, la condition*

$$\forall n \geq 0,\ \mathbb{E}(X_{n+1}|\mathcal{F}_n) = X_n$$

est équivalente à

$$\forall n > m,\ \mathbb{E}(X_n|\mathcal{F}_m) = X_m$$

Démonstration : On raisonne par récurrence.
Le résultat est vrai pour $n = m+1$. Si on suppose qu'il est vrai pour $n = m+k-1$, $k \geq 2$, alors par le théorème de conditionnement en cascade et la monotonie de l'espérance conditionnelle, on a

$$\mathbb{E}(Xn + k|\mathcal{F}_m) = \mathbb{E}(\mathbb{E}(X_{n+k}|\mathcal{F}_{m+k-1})|\mathcal{F}_m) = \mathbb{E}(X_{n+k-1}|\mathcal{F}_m)$$

□

Proposition 5.1.7. *Soit (X_n) une martingale. Alors*

$$\mathbb{E}(X_{n+1}) = \mathbb{E}(X_n) = \mathbb{E}(X_0)$$

Démonstration : On prend l'espérance dans la propriété de martingale, on a

$$\mathbb{E}(X_{n+1}) = \mathbb{E}(\mathbb{E}(X_{n+1}|\mathcal{F}_n)) = \mathbb{E}(X_n)$$

□

Proposition 5.1.8 (Martingales et Jensen)**.** *On va voir quelques propriétés des martingales associées avec de la convexité.*

1. *Si (X_n) est une martingale et $\varphi : \mathbb{R} \to \mathbb{R}$ est une fonction convexe telle que $\mathbb{E}(|\varphi(X_n)|) < +\infty$ pour tout $n \geq 0$ alors $Y_n = \varphi(X_n)$ est une sous-martingale.*
2. *Si (X_n) est une sous-martingale et $\varphi : \mathbb{R} \to \mathbb{R}$ est une fonction convexe et croissante telle que $\mathbb{E}(|\varphi(X_n)|) < +\infty$ pour tout $n \geq 0$ alors $Y_n = \varphi(X_n)$ est une sous-martingale.*

Démonstration : Dans le premier cas, Y_n est clairement $\mathcal{F}_n$-mesurable puisque X_n l'est et φ est convexe et mesurable. Puis, par l'inégalité de Jensen conditionnelle

$$\mathbb{E}(Y_{n+1}|\mathcal{F}_n) = \mathbb{E}(\varphi(X_{n+1})|\mathcal{F}_n) \geq \varphi\big(\mathbb{E}(X_{n+1}|\mathcal{F}_n)\big) = \varphi(X_n) = Y_n$$

D'où la sous-martingale.
Maintenant dans le deuxième cas, on a une sous-martingale et on utilise en plus la croissance de φ,

$$\mathbb{E}(Y_{n+1}|\mathcal{F}_n) = \mathbb{E}(\varphi(X_{n+1})|\mathcal{F}_n) \geq \varphi\big(\mathbb{E}(X_{n+1}|\mathcal{F}_n)\big) \geq \varphi(X_n) = Y_n$$

□

Corollaire 5.1.1. *Soit (X_n) une martingale avec $\mathbb{E}(|X_n|^p) < +\infty$ pour tout $n \geq 0$. Alors $(|X_n|^p)$ est une sous-martingale.*

Démonstration : On applique la proposition précédente avec $\varphi(x) = |x|^p$ convexe. □

Corollaire 5.1.2. *Soit $a \in \mathbb{R}$,*

1. *Soit (X_n) une sous-martingale. Alors $\big((X_n - a)^+\big)$ est une sous-martingale.*
2. *Soit (X_n) une sur-martingale. Alors, $\big(\min(X_n, a)\big)$ est une sur-martingale.*

Démonstration : Pour le premier point, $\varphi(x) = (x-a)^+$ est convexe croissante.
Pour le deuxième point, on applique le premier point à la sous-martingale $(-X_n)$ la fonction convexe croissante $\varphi(x) = \max(x, -a)$ pour avoir que $\varphi(-X_n)$ est une sous-martingale. Il s'ensuit que $-\varphi(-X_n) = -\max(-X_n, -a) = \min(X_n, a)$ forme une sur-martingale. □

On rappelle qu'une suite de variables aléatoires (H_n) est prévisible pour une filtration $(\mathcal{F}_n)$ si pour tout n, H_n est $\mathcal{F}_{n+1}$-mesurable.

Proposition 5.1.9. *Soit (X_n) une sous-martingale. Si (H_n) est une suite prévisible positive avec chaque H_n bornée, alors $(H.X)$ définie pour $(H.X)_0 = 0$ et*

$$(H.X)_n = \sum_{k=1}^{n} H_k(X_k - X_{k-1})$$

forme une sous-martingale. La même affirmation est vraie pour une sur-martingale ou pour une martingale sans la restriction de positivité $H_n \geq 0$ dans le cas d'une martingale.

Démonstration : On observe que $(H.X)_n$ est mesurable et que $(H.X)_n \in L^1$ car chaque $H_k(X_k - X_{k-1}) \in L^1$ puisque $X \in L^1$ et H est bornée. Ensuite, on a

$$(H.X)_n = (H.X)_{n-1} + H_n(X_n - X_{n-1})$$

Puis en utilisant la prévisibilité de H_n on a :

$$\mathbb{E}((H.X)_{n+1}|\mathcal{F}_n) = (H.X)_n + H_{n+1}\mathbb{E}(X_{n+1} - X_n|\mathcal{F}_n)$$

On conclut en observant que $\mathbb{E}(X_{n+1} - X_n|\mathcal{F}_n) \geq 0$ si (X_n) est une sous-martingale avec $H_n \geq 0$. On conclut de même avec $\mathbb{E}(X_{n+1} - X_n|\mathcal{F}_n) \leq 0$ si (X_n) est une sur-martingale et avec $\mathbb{E}(X_{n+1} - X_n|\mathcal{F}_n) = 0$ si c'est une martingale. □

Remarque 5.3 (Interprétation en termes financiers). On considère un actif risqué prenant la valeur X_n à la date n. Une suite prévisible (H_n) s'interprète dans ce contexte comme une stratégie d'investissement : Il s'agit de la quantité H_n d'actif risqué acheté à la date n. La valeur du portefeuille à la date n est alors

$$(H.X)_n = \sum_{i=1}^{n} H_i(X_i - X_{i-1})$$

En effet $(H.X)_n$ est la valeur $(H.X)_{n-1}$ à la date $n-1$ plus la valeur du nouvel actif H_nX_n moins le coût de l'achat H_nX_{n-1}.
On interprète également $(H.X)_n$ comme une intégrale stochastique (discrète) de (H_n) contre la suite (X_n).

La prévisibilité de H s'interprète alors de la façon suivante : chaque jour, les ordres d'achat sont passés le matin et les prix re-actualisés au cours de la journée. Ainsi, le jour n, la quantité H_n d'actif risqué est acheté à la valeur X_{n-1} du $(n-1)$-ième jour. La décision d'acheter est donc prise avec l'information dont on dispose à la date $n-1$, ie. les X_i, $i \leq n-1$ (il n'y a pas de délit d'initié). Cela justifie que la variable aléatoire H_n doit être $\mathcal{F}_{n-1}$ mesurable.

5.1.5 Martingale arrêtée

Etant donné un $(\mathcal{F}_n)$-temps d'arrêt et une suite $X = (X_n)$. La suite $X^T = (X_{T\wedge n})$ s'appelle la suite arrêtée.

Proposition 5.1.10. *Soit T un $(\mathcal{F}_n)$-temps d'arrêt.*

1. *Soit (X_n) une suite $(\mathcal{F}_n)$-adaptée. Alors $X^T = (X_{T\wedge n})$ est encore une suite $(\mathcal{F}_n)$-adaptée.*
2. *Soit (H_n) une suite $(\mathcal{F}_n)$-prévisible. Alors $H^T = (H_{T\wedge n})$ est encore une suite prévisible.*

Démonstration : Pour le premier point, on prend $B \in \mathcal{B}(\mathbb{R})$

$$\{X_n^T \in B\} = \left(\bigcup_{p=1}^{n} \{X_p \in B | T = p\} \right) \cup \{X_n \in B | T \geq n+1\} \in \mathcal{F}_n$$

puisque, pour $0 \leq p \leq n$, $\{X_p \in B\} \in \mathcal{F}_p \subseteq \mathcal{F}_n$, $\{T = p\} \in \mathcal{F}_p \subseteq \mathcal{F}_n$, $\{H_{n+1} \in B\} \in \mathcal{F}_n$, $\{T \geq n+1\} = \overline{\{T \leq n\}} \in \mathcal{F}_n$.
Pour le deuxième point, pour $B \in \mathcal{B}(\mathbb{R})$ on a

$$\{H_{n+1}^T \in B\} = \left(\bigcup_{p=0}^{n} \{H_p \in B | T = p\} \right) \cup \{H_{n+1} \in B | T \geq n+1\} \in \mathcal{F}_n$$

puisque, pour $0 \leq p \leq n$, $\{H_p \in B\} \in \mathcal{F}_{p-1} \subseteq \mathcal{F}_n$, $\{T = p\} \in \mathcal{F}_p \subseteq \mathcal{F}_n$, $\{H_{n+1} \in B\} \in \mathcal{F}_n$, $\{T \geq n+1\} = \overline{\{T \leq n\}} \in \mathcal{F}_n$. □

Définition 5.1.8 (Martingale arrêtée)**.** Si (X_n) est une $(\mathcal{F}_n)$-martingale et T est un $(\mathcal{F}_n)$-temps d'arrêt. On appelle martingale arrêtée la suite (X_n^T) avec $X_n^T = X_{T\wedge n}$. On introduit des notions analogues pour les sous-martingales ou sur-martingales.

On montre qu'une (sur/sous)-martingale arrêtée est une (sur/sous)-martingale.

Proposition 5.1.11 (Martingale arrêtée)**.** *Si T est un $(\mathcal{F}_n)$-temps d'arrêt et (X_n) une $(\mathcal{F}_n)$-martingale, sur ou sous-martingale. Alors, X^T est une $(\mathcal{F}_n)$-martingale, sur ou sous-martingale.*

Démonstration : La suite $H_n = \mathbb{1}_{T \geq n}$ est $(\mathcal{F}_n)$-prévisible. Dès lors

$$(H \cdot X)_n = \sum_{k=1}^{n} \mathbb{1}_{T \geq k}(X_k - X_{k-1}) = \sum_{k=1}^{T\wedge n} (X_k - X_{k-1}) = X_n^T - X_0$$

est une $(\mathcal{F}_n)$-martingale, sur ou sous-martingale selon ce qu'est X. Cela établit le résultat car la somme de martingale, sur ou sous-martingale est de même nature. □

Le théorème d'arrêt consiste à généraliser la propriété de martingale (sur/sous) à des dates $m \leq n$ données par des temps d'arrêts $S \leq T$. On commence par une version faible de cette propriété sur la constance (ou croissance/décroissance) des suites d'espérance.

D'abord, on donne une première forme du théorème d'arrêt pour des temps d'arrêt bornés.

Théorème 5.1.1 (Théorème d'arrêt - borné)**.** *Soit (X_n) une martingale et T un temps d'arrêt tel que $T \leq k$ presque sûrement pour un $k \in \mathbb{N}$ donné (i.e. T est borné). Alors $X_T \in L^1$ et*

$$\mathbb{E}(X_0) = \mathbb{E}(X_T) = \mathbb{E}(X_k)$$

De plus, si (X_n) est une sous-martingale, alors

$$\mathbb{E}(X_0) \leq \mathbb{E}(X_T) \leq \mathbb{E}(X_k)$$

et si (X_n) est une sur-martingale alors

$$\mathbb{E}(X_0) \geq \mathbb{E}(X_T) \geq \mathbb{E}(X_k)$$

Démonstration : Dans l'énoncé du théorème, comme on travaille souvent avec des martingales, on l'a énoncé dans cet ordre. Pour la démonstration, on commence par prouver le cas d'une sous-martingale. Le cas d'une sur-martingale s'obtiendra alors immédiatement en considérant $(-X_n)$. Le cas d'une martingale s'obtiendra alors immédiatement en combinant les deux approches.

Soit (X_n) une sous-martingale. Comme $0 \leq T \leq k$, on a $X_T = \sum_{i=0}^{k} X_i \mathbb{1}_{T=i}$ et il vient d'abord que $X_T \in L^1$ puisque $|X_T| \leq \sum_{i=0}^{k} |X_i|$.

Par les résultats précédents, $(X_{T \wedge n})$ est une sous-martingale. Ainsi, comme $0 \leq T \leq k$ ps, en utilisant la croissance des espérances pour la sous-martingale arrêtée X^T on a

$$\mathbb{E}(X_0) = \mathbb{E}(X_{T \wedge 0}) \leq \mathbb{E}(X_{T \wedge k}) = \mathbb{E}(X_T)$$

ce qui prouve la première inégalité. Pour la deuxième inégalité, on utilise la propriété de sous-martingale. Pour $0 \leq i \leq k$, on a $X_i \leq \mathbb{E}(X_k | \mathcal{F}_i)$ ps et comme $\{T = i\} \in \mathcal{F}_i$

$$\mathbb{E}(X_i \mathbb{1}_{T=i}) \leq \mathbb{E}(\mathbb{E}(X_k | \mathcal{F}_i) \mathbb{1}_{T=i}) = \mathbb{E}(X_k \mathbb{1}_{T=i})$$

Et donc en sommant

$$\mathbb{E}(X_T) = \mathbb{E}\left(\sum_{i=0}^{k} X_i \mathbb{1}_{T=i} \right) = \sum_{i=0}^{k} \mathbb{E}(X_T \mathbb{1}_{T=i}) \leq \sum_{i=0}^{k} \mathbb{E}(X_k \mathbb{1}_{T=i}) = \mathbb{E}(X_k)$$

Ce qui prouve le résultat pour une sous-martingale. Maintenant, comme dit précédemment, si (X_n) est une sur-martingale alors on applique ce résultat à la sous-martingale $(-X_n)$.
Si (X_n) est une martingale alors on a l'inégalité pour les sous-martingales pour (X_n) et pour $(-X_n)$, ce qui donne l'égalité voulue. □

On peut donner un contre-exemple dans le cas où le temps d'arrêt T n'est pas borné.

Exemple 5.9 (Contre-exemple). Soit (S_n) la marche aléatoire simple

$$S_n = \sum_{i=1}^{n} X_i$$

où les X_i sont i.i.d. de loi de Rademacher $_Pro(X_1 = 1) = \mathbb{P}(X_1 = -1) = \frac{1}{2}$ et $S_0 = 0$. Il s'agit d'une martingale pour la filtration $\mathcal{F}_n = \sigma(X_1, ..., X_n)$, $n \geq 1$.
On note $T = \inf\{n \geq 0 | S_n = -1\}$. Il s'agit bien d'un temps d'arrêt. Alors $\mathbb{E}(S_0) = 0 > -1 = \mathbb{E}(S_T)$.
On note que T n'est pas borné puisque

$$\{T \leq n\} \supseteq \{X_1 = 1, ...X_n = 1\}$$

et donc

$$\mathbb{P}(T \geq n) \geq \mathbb{P}(X_1 = 1, ..., X_n = 1) = \frac{1}{2^n}$$

Si bien que la première inégalité n'est pas automatique si T n'est pas borné.

Théorème 5.1.2. *Soit (X_n) une martingale et T un temps d'arrêt. Sous chacune des conditions suivantes on a $X^T \in L^1$ et*

$$\mathbb{E}(X_0) = \mathbb{E}(X_T)$$

1. *T est borné (i.e. il existe $C > 0$ tel que $T \leq C$ presque sûrement).*
2. *La suite X est bornée (i.e. il existe $K > 0$ tel que $|X_n| \leq K$ pour tout n ps) et T est fini presque sûrement.*
3. *$\mathbb{E}(T) < \infty$ et il existe $K > 0$ tel que $|X_{n+1} - X_n| \leq K$ ps pour tout $n \geq 0$.*

De plus, si (X_n) est une sous-martingale alors on a $\mathbb{E}(X_0) \leq \mathbb{E}(X_T)$ sous les mêmes conditions.
Si (X_n) est une sur-martingale alors on a $\mathbb{E}(X_0) \geq \mathbb{E}(X_T)$ sous les mêmes conditions ou encore sous $X_n \geq 0$ et T fini ps.

Démonstration : On reprend le même schéma que la dernière démonstration. Supposons que (X_n) soit une sous-martingale.
Le premier point découle du théorème précédent.
Pour le deuxième point, comme $T \wedge n$ est un temps d'arrêt borné alors on applique le premier point

$$\mathbb{E}(X_0) \leq \mathbb{E}(X_{T \wedge n})$$

Maintenant, on a $X_{T \wedge n} \to X_T$ (car $T < \infty$) et

$$X_{T \wedge n} = \sum_{i=0}^{\infty} X_{T \wedge n} \mathbb{1}_{T=i} = \sum_{i=0}^{\infty} X_{n \wedge i} \mathbb{1}_{T=i}$$

et,

$$|X_{T \wedge n}| \leq \sum_{i=0}^{\infty} |X_{n \wedge i}| \mathbb{1}_{T=i} \leq K \sum_{i=0}^{\infty} \mathbb{1}_{T=i} = K$$

comme $X_{T\wedge n} \to X_T$, on a aussi $|X_T| \leq K$. On a donc $X_T \in L^1$. Puis, par le théorème de convergence dominée

$$\mathbb{E}(X_T) = \lim_{n\to\infty} \mathbb{E}(X_{T\wedge n}) \geq \mathbb{E}(X_0)$$

Finalement pour le troisième point, on a toujours (en appliquant le premier à $T \wedge n$)

$$\mathbb{E}(X_0) \leq \mathbb{E}(X_{T\wedge n})$$

avec $X_{T\wedge n} \to X_T$ ps ($T < \infty$ ps). On peut écrire

$$X_T = X_0 + \sum_{k=1}^{T\wedge n} (X_k - X_{k-1})$$

et

$$|X_T| \leq |X_0| + \sum_{k=1}^{T\wedge n} |X_k - X_{k-1}| \leq |X_0| + KT \in L^1$$

car $\mathbb{E}(T) < \infty$. Le théorème de convergence dominée s'applique et donne

$$\mathbb{E}(X_T) = \lim_{n\to\infty} \mathbb{E}(X_{T\wedge n}) \geq \mathbb{E}(X_0)$$

De plus,

$$\mathbb{E}(|X_T|) = \lim_{n\to\infty} \mathbb{E}(|X_{T\wedge n}|) \leq \mathbb{E}(|X_0|) + K\mathbb{E}(|T|) < \infty$$

soit $X_T \in L^1$.
On a traité le cas d'une sous-martingale, maintenant on explore les deux autres options.
Si (X_n) est une martingale alors en particulier on a égalité

$$\mathbb{E}(X_0) = \mathbb{E}(X_T)$$

et les passages à la limite préservent évidemment les égalités. De plus, comme (X_n) est en particulier une sous-martingale alors on a toujours $X_T \in L^1$.
Pour finir, si (X_n) est une sur-martingale alors les 3 points étant insensibles aux changements de signes, on applique le cas de la sous-martingales à $(-X_n)$ pour obtenir $\mathbb{E}(X_0) \geq \mathbb{E}(X_T)$. Puis, sous le quatrième point, partant de $\mathbb{E}(X_{T\wedge n}) \leq \mathbb{E}(X_0)$ pour la sur-martingale arrêtée (X_n^T), le lemme de Fatou donne

$$\mathbb{E}(X_T) = \mathbb{E}\big(\lim_{n\to\infty} X_{T\wedge n}\big) = \mathbb{E}\big(\liminf_{n\to\infty} X_{T\wedge n}\big) \leq \liminf_{n\to\infty} \mathbb{E}(X_{T\wedge n}) \leq \mathbb{E}(X_0)$$

en particulier comme $X_T \geq 0$ alors on a aussi $X_T \in L^1$ puisque $\mathbb{E}(X_0) < \infty$.□

5.1.6 Décomposition de Doob

Théorème 5.1.3 (Décomposition de Doob)**.** *Toute $(\mathcal{F}_n)$-sous-martingale (X_n) se décompose de façon (presque sûrement) unique sous la forme*

$$X_n = M_n + A_n$$

où (M_n) est une $(\mathcal{F}_n)$-martingale (fluctuations aléatoires) et (A_n) est une suite croissante $(\mathcal{F}_n)$-prévisible ("drift" prévisible i.e. tendance déterministe) avec $A_0 = 0$ et donnée par

$$A_n = \sum_{k=1}^{n} \big(\mathbb{E}(X_k|\mathcal{F}_{k-1}) - X_{k-1}\big)$$

Démonstration : Après. □

Ainsi, une sous-martingale est enfait une martingale plus un processus croissant A_n. Cela simplifie grandement l'étude de sa convergence ou de ses propriétés asymptotiques par exemple.
Aussi, les propriétés sur les martingales (comme les convergence p.s. ou L^p) s'étendent facilement aux sur/sous-martingales grâce à la décomposition, en contrôlant le processus prévisible A_n.

Remarque 5.4. La décomposition est vraie pour toute suite (X_n) adaptée e L^1 avec (M_n) martingale et (A_n) prévisible partant de $A_0 = 0$. En fait, (X_n) est une sous-martingale si et seulement si (A_n) est croissante.

Soit (X_n) une martingale de carrée intégrable et nulle en 0. Comme (X_n^2) est une sous-martingale alors on a la décomposition de Doob suivante

$$X_n^2 = M_n + A_n$$

Définition 5.1.9 (Compensateur)**.** On note $< X, X >$ le processus A croissant prévisible dans la décomposition de Doob de X^2 et on l'appelle compensateur de la martingale $X \in L^2$.

$$A_n = \langle X, X\rangle_n = \sum_{k=1}^{n} \big(\mathbb{E}(X_k^2|\mathcal{F}_{k-1}) - X_{k-1}^2\big)$$

$$= \sum_{k=1}^{n} \mathbb{E}\big((X_k - X_{k-1})^2|\mathcal{F}_{k-1}\big)$$

En effet, la deuxième formulation vient du fait que pour une martingale telle que $\mathbb{E}(X_n^2) < \infty$ alors pour $n > k$

$$\begin{aligned}\mathbb{E}\big((X_n - X_k)^2|\mathcal{F}_k\big) &= \mathbb{E}(X_n^2|\mathcal{F}_k) - 2X_k\mathbb{E}(X_n|\mathcal{F}_k) + X_k^2 \\ &= \mathbb{E}(X_n^2|\mathcal{F}_k) - X_k^2\end{aligned}$$

On appelle cette expression, la formule de la variance conditionnelle.

Dans l'expression que l'on a, $\langle X, X\rangle_n$ apparait comme la variance jusqu'à la date n et $\langle X, X\rangle_\infty$ (qui existe toujours par croissance, quitte à avoir ∞) est la variance totale de toute la suite (X_n).

Le comportement L^2 d'une martingale de carré intégrable peut se lire sur son compensateur.

Proposition 5.1.12 (Martingale bornée dans L^2 et compensateur)**.** *Soit (X_n) une martingale carré intégrable. Alors elle est bornée dans L^2 si et seulement si son compensateur vérifie*

$$\mathbb{E}(\langle X, X\rangle_\infty) < \infty$$

Démonstration : Comme $(X_n^2 - \langle X, X\rangle_n)$ est une martingale alors

$$\mathbb{E}(X_n^2 - \langle X, X\rangle_n) = \mathbb{E}(X_0^2 - \langle X, X\rangle) = \mathbb{E}(X_0^2)$$

ie,

$$\mathbb{E}(\langle X, X\rangle_n) = \mathbb{E}(X_n^2) - \mathbb{E}(X_0^2),\ \forall n \geq 0$$

Comme $\langle X, X\rangle_n$ est une suite croissante, le théorème de convergence monotone donne alors

$$\mathbb{E}\big(\langle X, X\rangle_\infty\big) = \lim_{n\to\infty} \mathbb{E}\big(\langle X, X\rangle_n\big) = \sup_n \mathbb{E}\big(\langle X, X\rangle_n\big) = \sup_n \mathbb{E}(X_n^2) - \mathbb{E}(X_0^2)$$

□

Proposition 5.1.13. *Soit (X_n) une martingale de carré intégrable, nulle en 0 et T un temps d'arrêt. Alors*

$$\langle X^T, X^T\rangle = \langle X, X\rangle^T,\ ps$$

ie. le crochet de la martingale arrêtée est le crochet arrêté de la martingale.

Démonstration : On a

$$(X^T)^2 = M^T + \langle X, X\rangle^T$$

est une sous-martingale, M^T est une martingale et $\langle X, X\rangle^T$ est une suite croissante et prévisible. L'unicité (presque sûrement) de la décomposition de Doob de X^T exige

$$\langle X^T, X^T\rangle = \langle X, X\rangle^T,\ ps$$

□

Exemple 5.10 (Somme de variables aléatoires iid). Soit (X_n) une suite de variables aléatoires iid centrées, de carrés intégrables avec $\mathrm{Var}(X_1) = \sigma^2$. Alors S_n est une martingale L^2 de compensateur $n\sigma^2$. En effet, en utilisant $X_k \perp\!\!\!\perp \mathcal{F}_{k-1} = \sigma(X_1, ..., X_{k-1})$ on a

$$\langle S, S\rangle_n = \sum_{k=1}^n \mathbb{E}\big((S_k - S_{k-1})^2 | \mathcal{F}_{k-1}\big) = \sum_{k=1}^n \mathbb{E}(X_k^2 | \mathcal{F}_{k-1}) = \sum_{k=1}^n \mathbb{E}(X_k^2) = n\sigma^2$$

5.2 Convergence de Martingales

Dans ce chapitre on va maintenant étudier les limites de martingales. On commence par les outils clef que sont les inégalités pour les martingales (dues à Doob). On donne ensuite des résultats de convergence ps puis en norme L^1 et en norme L^p. On donne ensuite le résultat fondamental qu'est le théorème d'arrêt, qui généralise la propriété de martingales aux dates données par des temps d'arrêt.

Dans la suite on considère un espace de probabilité filtré. Par défaut, les (sur/sous)-martingales et temps d'arrêt seront considérés par rapport à cette filtration $(\mathcal{F}_n)$.

De plus, étant donné une suite (X_n) on notera

$$\overline{X_n} = \max_{0\leq k\leq n} X_k$$

5.2.1 Inégalités de martingales : Inégalité maximale de Doob

Théorème 5.2.1 (Inégalité maximale de Doob)**.**

1. *Soit (X_n) une sous-martingale et $x > 0$. Alors*

$$\mathbb{P}(\overline{X_n} \geq x) \leq \frac{\mathbb{E}\big(X_n \mathbb{1}_{\overline{X_n} \geq x}\big)}{x} \leq \frac{\mathbb{E}\big(X_n^+\big)}{x} \leq \frac{\mathbb{E}\big(|X_n|\big)}{x}$$

2. *Soit (X_n) une sur-martingale et $x > 0$. Alors*

$$\mathbb{P}(\overline{X_n} \geq x) \leq \frac{\mathbb{E}\big(|X_0|\big) + \mathbb{E}\big(|X_n|\big)}{x}$$

3. *Soit (X_n) une sur-martingale* ***positive*** *et $x > 0$. Alors*

$$\mathbb{P}(\overline{X_n} \geq x) \leq \frac{\mathbb{E}(X_0)}{x}$$

4. *Soit (X_n) une martingale, sous-martingale ou sur-martingale et $x > 0$, on a*

$$\mathbb{P}\big(\max_{0 \leq k \leq n} |X_k| \geq x\big) \leq \frac{\mathbb{E}(|X_0|) + 2\mathbb{E}(|X_n|)}{x}$$

5. *Pour une martingale (X_n), on peut améliorer l'inégalité précédente en*

$$\mathbb{P}\big(\max_{0 \leq k \leq n} |X_k| \geq x\big) \leq \frac{\mathbb{E}(|X_n|)}{x}, \ x > 0$$

Démonstration : Pour le premier point, les deux inégalités de droite sont immédiates. On prouve alors celle de gauche. Notons $A = \{\overline{X_n} \geq x\}$ et, on pose $T = S \wedge n$ avec

$$S = \inf\{k \geq 0 | X_k \geq x\}$$

Comme $T \leq n$ on a $\mathbb{E}(X_T) \leq \mathbb{E}(X_n)$ ou encore

$$\mathbb{E}(X_T \mathbb{1}_A) + \mathbb{E}(X_T \mathbb{1}_{\overline{A}}) = \mathbb{E}(X_T) \leq \mathbb{E}(X_n) = \mathbb{E}(X_n \mathbb{1}_A) + \mathbb{E}(X_n \mathbb{1}_{\overline{A}})$$

Sur l'évènement $\overline{A} = \{\overline{X_n} < x\}$ on a $S > n$ et donc $T = n$ et $X_T = X_n$. Il vient alors que $\mathbb{E}(X_T \mathbb{1}_{\overline{A}}) = \mathbb{E}(X_n \mathbb{1}_{\overline{A}})$ et donc l'inégalité du dessus s'écrit $\mathbb{E}(X_T \mathbb{1}_A) \leq \mathbb{E}(X_n \mathbb{1}_A)$. De plus sur A, on a $S \leq n$ donc $T = S$ et par définition de S : $X_T = X_S \geq x$. Finalement, il vient

$$x\mathbb{P}(A) = \mathbb{E}(x \mathbb{1}_A) \leq \mathbb{E}(X_T \mathbb{1}_A) \leq \mathbb{E}(X_n \mathbb{1}_A)$$

ce qui prouve le premier point car $x > 0$.

Pour les deuxième et troisième points, on adapte la preuve du premier au cas d'une sur-martingale X. Pour $T \leq n$, alors $\mathbb{E}(X_T) \leq \mathbb{E}(X_0)$, soit

$$\mathbb{E}(X_T \mathbb{1}_A) + \mathbb{E}(X_T \mathbb{1}_{\overline{A}}) = \mathbb{E}(X_T) \leq \mathbb{E}(X_0)$$

comme précédemment $\mathbb{E}(X_T \mathbb{1}_{\overline{A}}) = \mathbb{E}(X_n \mathbb{1}_{\overline{A}})$ et $\mathbb{E}(X_T \mathbb{1}_A) \geq x\mathbb{P}(A)$ et l'inégalité du dessus donne les deux points car en général

$$x\mathbb{P}(A) \leq \mathbb{E}(X_0) - \mathbb{E}(X_n \mathbb{1}_{\overline{A}}) \leq \mathbb{E}(|X_0|) + \mathbb{E}(|X_n|)$$

d'où le deuxième point, et lorsque la sur-martingale est positive, ie $X_n \geq 0$,

$$x\mathbb{P}(A) \leq \mathbb{E}(X_0) - \mathbb{E}(X_n \mathbb{1}_{\overline{A}}) \leq \mathbb{E}(X_0)$$

d'où le troisième point.

Pour le quatrième point, comme $\{\max_{k\leq n} |X_k| \geq x\} \subseteq \{\max_{k\leq n} X_k \geq x\} \cup \{\max_{k\leq n}(-X_k) \geq x\}$, on a

$$\mathbb{P}(\max_{0\leq k\leq n} |X_k| \geq x) \leq \mathbb{P}(\max_{0\leq k\leq n} X_k \geq x) + \mathbb{P}(\max_{0\leq k\leq n} (-X_k) \geq x)$$

et (X_n), $(-X_n)$ sont des sous-martingales et sur-martingales (ou l'inverse) ainsi on majore chaque terme par les premier et deuxième points (pour respectivement la sur et sous-martingale), ce qui donne le résultat.

Pour le cinquième et dernier point, lorsque (X_n) est une martingale, on peut appliquer le point 1 à la sous-martingale positive $(|X_n|)$ et avoir le résultat.□

Corollaire 5.2.1 (Inégalité maximale de Kolmogorov)**.** *Soit (X_n) des variables aléatoires indépendantes centrées et de variances finies. On pose $S_n = X_1 + ... + X_n$. Alors, pour $x > 0$ on a*

$$\mathbb{P}\left(\max_{1\leq k\leq n} |S_k \geq x\right) \leq \frac{Var(S_n)}{x^2}$$

Démonstration : On a vu que (S_n) est une martingale pour la filtration canonique engendrée par la suite (X_n). Ainsi, $Y_n = S_n^2$ est une sous-martingale à laquelle on applique l'inégalité maximale de Doob avec $u = x^2$. Puisque $\mathbb{E}(S_n^2) = \mathrm{Var}(S_n)$ (car les variables aléatoires sont centrées), on obtient l'inégalité voulue. □

Maintenant, dans le contexte de ce corollaire, l'inégalité de Tchebychev nous donne

$$\mathbb{P}(|S_k| \geq x) \leq \frac{\mathrm{Var}(S_k)}{x^2} \leq \frac{\mathrm{Var}(S_n)}{x^2}$$

car

$$\mathrm{Var}(S_k) = \sum_{i=1}^{k} \mathbb{E}(X_i^2) \leq \sum_{i=1}^{n} \mathbb{E}(X_i^2) = \mathrm{Var}(S_n)$$

On a donc

$$\max_{1\leq k\leq n} \mathbb{P}(|S_k| \geq x) \leq \frac{\mathrm{Var}(S_n)}{x^2}$$

Comme

$$\max_{1\leq k\leq n} \mathbb{P}(|S_k| \geq x) \leq \mathbb{P}\left(\bigcup_{k=1}^{n} \{|S_k| \geq x\}\right) = \mathbb{P}\left(\max_{1\leq k\leq n} |S_k| \geq x\right)$$

Si bien que l'inégalité maximale de Kolmogorov est meilleure que l'inégalité de Tchebychev dans ce contexte.

5.2.2 Inégalités de martingales : Inégalité de moments de Doob

Théorème 5.2.2 (Inégalité de moments pour sous-martingale)**.**

5.2.3 Inégalités de martingales : Nombre de montées

5.2.4 Convergence presque-sûre de martingales

5.2.5 Uniforme intégrabilité

La notion d'uniforme intégrabilité qu'on intriduit maintenant sera utile pour étudier la convergence dans L^1 de martingales.

Définition 5.2.1 (Uniforme intégrabilité)**.** Une suite de variables aléatoires intégrables (X_n) est dite uniformément intégrable (UI) si

$$\lim_{c\to\infty} \sup_{n\geq 0} \mathbb{E}\big(|X_n|\mathbb{1}_{|X_n|>c}\big) = 0$$

5.2.6 Convergence L^1 et martingales fermées

5.2.7 Convergence L^p de martingales pour $p > 1$

5.2.8 Martingales carré-intégrables

5.2.9 Théorème d'arrêt

6 Statistiques (M1)

Sources : **[1]** Statistiques Mathématiques, Clément Levrard, Université de Rennes - ENS Rennes, 2023-2024.
[2] Statistiques Mathématiques, Benoît Cadre - Céline Vial, Editions Ellipses, 2012.

6.1 Modèles statistique

Jusqu'ici dans le parcours Mathématiques, on a fait très peu de statistiques et beaucoup de probabilités. Quelle est la différence entre le point de vue probabiliste et le point de vue statistique ? La réponse à cette question réside principalement dans la nature des questions auxquelles ils cherchent à répondre et les méthodes utilisées.

Le point de vue probabiliste ne cherche pas à observer la variable aléatoire X mais va plutôt chercher à la modéliser. Ce point de vue se concentre donc principalement sur la modélisation des phénomènes aléatoires et à la prédiction des évènements futurs. Il se concentre sur les propriétés des distributions de probabilité et les lois qui régissent les évènements aléatoires.

Le point de vue statistique se concentre sur l'analyse des données observées pour faire des inférences sur les populations ou les processus sous-jacents. Il s'agit de tirer des conclusions à partir de données échantillonnées.

6.1.1 Modèles statistique

Définition 6.1.1 (Modèle statistique)**.** Un modèle statistique est un triplet $(\mathcal{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ composé des éléments suivants

- $\mathcal{X}$: Espace des observations ;
- $\mathcal{A}$: Tribu sur $\mathcal{X}$, pour donner un sens à ce qui est observable ou non. $(\mathcal{X}, \mathcal{A})$ est un espace mesurable ;
- $(\mathbb{P}_\theta)_{\theta\in\Theta}$: Famille de lois sur $\mathcal{X}$ indexée par $\theta \in \Theta$ (où Θ, souvent $\subseteq \mathbb{R}^p$ est appelé espace des paramètres).

Il est implicitement supposé qu'on observe X de loi $\mathbb{P}_\theta$, pour θ inconnu, et qu'on cherche à estimer une quantité $q(\theta)$ à partir de ces observations. On notera (pour la suite des cours dans ce polycopié) qu'en apprentissage θ est remplacé par une fonction f et $\mathbb{P}_\theta$ par une loi $\mathbb{P}$ inconnue sur (X, Y).

On note que par convention on confondra souvent X de loi $\mathbb{P}_\theta$ et $x = X(\omega)$ (observation). Cette convention se justifie par le fait qu'on s'intéresse aux observations uniquement au travers de leur loi pour essayer d'avoir θ.

On remarque aussi que par convention, dans les modèles "classiques" où l'espace des observations et la tribu associée sont évidents, on pourra omettre de les mentionner. Par exemple le modèle $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (\mathcal{N}(\theta, \sigma_0^2)^{\oplus n})_{\theta\in\Theta})$, i.e. un tirage de n gaussiennes inépendantes de variance connue et de moyenne inconnue, est souvent abrégé en $(\mathcal{N}(\theta, \sigma_0^2)^{\oplus n})_{\theta\in\Theta}$.

Dernière chose, si $(\mathcal{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ est un modèle et si $g : \Theta \to \Theta$ est une bijection, alors $(\mathcal{X}, \mathcal{A}, (\mathbb{P}_{g(\theta)})_{\theta\in\Theta})$ est un modèle qui est équivalent d'un point de vue statistique. Si on anticipe légèrement, chercher à estimer $q(\theta)$ dans le modèle $\mathbb{P}_\theta$ revient à chercher à estimer $q \circ g(\theta)$ dans le modèle $\mathbb{P}_{g(\theta)}$.

Traditionnellement, on distingue deux grands types de modèles.

Modèle paramétriques :

C'est le cas où l'espace des paramètres Θ est un sous-ensemble de $\mathbb{R}^d$ ou plus généralement est de dimension finie. Quelques exemples de modèles paramétriques :

- Régression linéaire : Suppose que la relation entre les variables dépendantes et indépendantes est linéaire.
- Distribution normale : Suppose que les données suivent une distribution gaussienne, caractérisée par une moyenne μ et une variance σ^2.
- Modèles de Poisson : Utilisés pour modéliser des données de comptage, avec un paramètre de taux λ.

Avec ce genre de modèles, on a certes moins de paramètres à estimer ce qui présente un certain avantage en terme d'efficacité de calcuk et de stockage de données, cependant souvent on observe une rigidité de ce modèle car les hypothèses peuvent être trop restrictives ce qui les rendent difficiles à adapter à des données réelles.

Modèle non-paramétrique :

C'est le cas où l'espace des paramètres Θ n'est pas de dimension finie. Ces modèles ne font pas d'hypothèses spécifiques sur la forme de la distribution de données. Ils sont donc plus flexibles et peuvent donc s'avérer moins sensibles aux potentielles hypothèses incorrectes sur la forme de la distribution. Cependant ils peuvent être plus complexes à interpréter et à calculer.

Exemple 6.1 (Modélisation de la pollution atmosphérique à Paris). Si on dispose de mesures quotidiennes de concentration en particules fines à Paris, que l'on suppose les mesures indépendantes et de même loi, et que cette loi commune à une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$, le modèle statistique correspond à

- $\mathcal{X} = \mathbb{R}^n$;
- $\mathcal{A} = \mathcal{B}(\mathbb{R}^n)$;
- $\Theta = \{\text{densité sur } \mathbb{R}\}$. Pour $\theta \in \Theta$, $\mathbb{P}_\theta = (\theta d\lambda)^{\oplus n}$

Si maintenant Paris est modélisé par un carré $[0, T]^2$, et que pour chaque journée i à 12h on dispose du relevé de concentration en particules fines à chaque endroit $X^{(i)} : [0, T]^2 \to \mathbb{R}^+$. Une modélisation frustre serait de considérer que les relevés journaliers sont indépendants et de même loi que celle de $X : t \mapsto f(t) + \varepsilon_t$, où f est continue et ε_t est un bruit gaussien de fonction de covariance connue. Dans ce cas le modèle est

- $\mathcal{X} = \mathcal{C}([0,T]^2, \mathbb{R})^n$;
- $\mathcal{A} = \mathcal{B}(\mathcal{C}([0,T]^2, \mathbb{R}^+))^{\oplus n}$ (plus petite tribu qui rend les applications coordonnées mesurables) ;
- $\Theta = \mathcal{C}([0,T]^2, \mathbb{R}^+)$. Pour $\theta \in \Theta$, $\mathbb{P}_\theta = (\mathcal{L}(\theta + \varepsilon))^{\oplus n}$.

En pratique les mesures observées ne sont pas de type fonctionnel mais plutôt sur un grillage de $[0,T]^2$. Le modèle de bruit gaussien nous ramène dans ce cas à un modèle gaussien paramétrique (à structure de covariance connue) de très grande dimension (nombre de pixels).

La pluaprt du temps on considérera des modèles correspondant à la réalisation de n variables aléatoires indépendantes et de même loi. On parle alors de n-échantillon, qui correspondent à des modèles du type $(\mathcal{X}^n, \mathcal{A}\oplus n, (\mathbb{P}_\theta^{\oplus n})_{\theta\in\Theta})$. Dans ce type de modèle, on notera $X_{1:n}$ un vecteur aléatoire de loi $\mathbb{P}_\theta^{\oplus n}$.

6.1.2 Estimation ponctuelle

Définition 6.1.2 (Statistique)**.** Pour un modèle $(\mathcal{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$, une statistique est une fonction mesurable sur $(\mathcal{X}, \mathcal{A})$.

En somme, une statistique est une fonction des observations qui ne peut prendre θ comme argument. Sa dépendance à θ ne se fait qu'au travers de $\mathbb{P}_\theta$.

$$\mathbb{E}_\theta(f(X)) = \int_{\mathcal{X}} f(u)\mathbb{P}_\theta(du)$$

Si le but est de deviner θ à partir des observations alors il est naturel de considérer des statistiques à valeurs dans Θ :

Définition 6.1.3 (Estimateur)**.** Dans le modèle $(\mathcal{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$, un estimateur de $q(\theta)$, où $q : \Theta \to \mathcal{Y}$ est juste une statistique à valeurs dans $\mathcal{Y}$.

Un estimateur est alors une application mesurable

$$T_n : \mathcal{X}^n \to \mathbb{R}^k$$

qui a comme but d'estimer le paramètre d'intérêt $q(\theta)$. On jugera sa qualité via une fonction de perte

$$l : (q(\theta), T_n) \in \mathbb{R}^k \times \mathbb{R}^k \to l(q(\theta), T_n) \in \mathbb{R}_+$$

En statistique paramétrique, on prendra toujours $\mathcal{Y} = \mathbb{R}^k$.

Exemple 6.2. Dans le modèle $(\mathcal{N}(\theta, \sigma_0^2)^{\oplus n})_{\theta\in\mathbb{R}}$, un estimateur standard de θ est la moyenne empirique

$$\hat{\theta} = \overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i$$

Exemple 6.3. Dans le modèle $(\mathcal{U}(]0,\theta[)^{\oplus n})_{\theta>0}$, deux estimateurs raisonnables de θ sont

$$\hat{\theta} = 2 * \overline{X_n}$$

$$\hat{\theta} = \max_{i=1,\dots,n} X_i$$

Risque quadratique :

Une manière d'évaluer la qualité d'estimation ponctuelle est de considérer le risque quadratique de l'estimateur T.

Définition 6.1.4 (Risque quadratique)**.** Dans le modèle $(\mathcal{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$, le risque quadratique de T est

$$R_T(\theta) = \mathbb{E}_\theta ||T(X) - q(\theta)||^2$$

On peut décomposer le risque quadratique de la sorte

$$R_T(\theta) = ||\mathbb{E}_\theta(T(X)) - q(\theta)||^2 + \mathrm{Var}_\theta(T(X))$$

- $\mathbb{E}_\theta(T(X)) - q(\theta)$: Biais de l'estimateur T ;
- $\mathrm{Var}_\theta(T(X)) = \mathbb{E}_\theta\big(||T(X) - \mathbb{E}_\theta(T(X))||^2\big)$: Variance de l'estimateur T sous $\mathbb{P}_\theta$.

On voit directement qu'on préférera un estimateur avec un biais faible, voir sans biais. Un estimateur sans biais est un estimateur qui en moyenne donne la bonne réponse.

Exemple 6.4. Dans le modèle $(\mathcal{U}(]0,\theta[)^{\oplus n})_{\theta\in\Theta}$

- $T_1(X_{1:n}) = \hat{\theta} = 2 * \overline{X_n}$ est sans biais.

$$R_{T_1}(\theta) = \frac{\theta^2}{3n}$$

- $T_2(X_{1:n}) = \max_{i=1,\dots,n} X_i$ est biaisé.

$$R_{T_2}(\theta) = \frac{\theta^2}{(n+1)(n+2)}$$

Dans l'exemple i-dessus, le risque de T_2 est sensiblement meilleur que celui de T_1. Ainsi, bien que le caractère sans biais soit souhaitable, on observe qu'il ne garantit pas l'optimalité.

Comportements asymptotiques souhaitables :

Quand on parle de "convergence" d'estimateurs, on se place dans le modèle $(\mathcal{X}^n, \mathcal{A}^{\oplus n}, (\mathbb{P}_\theta^{\oplus n})_{\theta\in\Theta})$, i.e. on observe un n-échantillon indépendants et de même loi.
Une propriété minimale des estimateurs est que lorsque l'information disponible croît, l'estimateur converge vers la valeur souhaitée.

Définition 6.1.5 (Consistance)**.** Un(e) (suite d') estimateur(s) T de $q(\theta) \in \mathbb{R}^k$ est dit consistant si

$$\forall\theta \in \Theta,\ T(X_{1:n}) \to^{\mathbb{P}} q(\theta)$$

Si la convergence a lieu presque-sûrement, on parle de consistance forte.

Prouver une consistance peut se faire via la LGN, ou en utilisant la convergence du risque quadratique vers 0.

Définition 6.1.6 (Normalité asymptotique)**.** Dans le cas où $q(\theta) \in \mathbb{R}$, un estimateur T est dit asymptotiquement normal en θ s'il existe une suite r_n positive et $\sigma_\theta^2 > 0$ tels que

$$r_n(T(X_{1:n}) - q(\theta)) \to \mathcal{N}(0, \sigma_\theta^2)$$

La normalité asymptotique en θ est la convergence en loi de l'estimateur renormalisé vers une loi normale non-dégénérée.
On peut étendre la définition en dimension supérieure en rajoutant l'hypothèse d'une matrice de covariance nulle.

La normalité asymptotique n'est pas intéressante en elle-même. L'idée est de chercher le comportement asymptotique de la statistique recentrée pour pouvoir en déduire ultérieurement des garanties en terme de risque asymptotique ou d'intervalle de confiance.

Notons que le TCL indique que le comportement asymptotique normal est relativement fréquent. En effet, le théorème nous dit que lorsque l'on prend des échantillons de grandes tailles, la distribution de la moyenne de ces échantillons tendra à être normale, indépendamment de la distribution des données originales.

Exemple 6.5. On reprend le modèle $(\mathcal{U}(]0,\theta[)^{\oplus n})$ et les 2 estimateurs précédents associés à ce modèle.

- $\mathbb{E}_\theta(X_1^2) < \infty$, le TCL donne, avec $T_1(X_{1:n}) = 2\overline{X_n}$ et $R_{T_1}(\theta) = \dfrac{\theta^2}{3n}$,

$$\sqrt{n}(T_1(X_{1:n}) - \theta) \to \mathcal{N}(0, \frac{\theta^2}{3})$$

- Pour T_2, si on prend $t > 0$

$$\mathbb{P}_\theta(n(\theta - T_2(X_{1:n})) > t) = \left(1 - \frac{t}{\theta}\right)^n \mathbb{1}_{t \leq n\theta} \to e^{-\frac{t}{\theta}}$$

  Si bien que,

$$n(\theta - T_2(X_{1:n})) \to \mathcal{E}(\theta^{-1})$$

  Il y a un comportement asymptotique mais il n'est pas normale.

Etudions maintenant comment prouver une normalité asymptotique à l'aide du TCL et de 2 nouveaux outils : le Lemme de Slutsky et la Δ-méthode.

Théorème 6.1.1 (Lemme de Slutsky)**.** *Soient X_n et Y_n deux suites de vecteurs aléatoires convergeant en loi respectivement vers X et y (vecteur aléatoire constant valant y p.s.).*
Alors, $Y_n \to^{\mathbb{P}} y$ et

$$(X_n, Y_n) \to^{\mathcal{L}} (X, y)$$

Démonstration : Premièrement, on rappelle qu'un couple (X_n, Y_n) converge en loi vers un couple (X, Y) si pour toute fonction continue et bornée $\varphi : \mathbb{R}^2 \to \mathbb{R}$, on a

$$\mathbb{E}(\varphi(X_n, Y_n)) \to \mathbb{E}(\varphi(X, Y))$$

De plus, modulo l'introduction de la fonction caractéristique d'un couple aléatoire (X, Y) via

$$\forall s, t \in \mathbb{R},\ \Phi_{(X,Y)}(s, t) = \mathbb{E}\left(e^{i(sX+tY)}\right)$$

Le théorème de Levy est toujours valide. C'est-à-dire que (X_n, Y_n) converge en loi vers (X, Y) si et seulement si

$$\forall s, t \in \mathbb{R},\ \Phi_{(X,Y_n)}(s, t) \to \Phi_{(X,Y)}(s, t)$$

On a donc résumé ce que l'on doit prouver, i.e. que pour tout couple de réels $(s, t) \in \mathbb{R}^2$,

$$\mathbb{E}\left(e^{i(sX_n+tY_n)}\right) \to \mathbb{E}\left(e^{i(sX+ty)}\right)$$

Pour cela, on part de

$$e^{i(sX_n+tY_n)} - e^{i(sX+ty)} = e^{isX_n}(e^{itY_n} - e^{ity}) + e^{ity}(e^{isX_n} - e^{isX})$$

D'où,

$$|\Phi_{(X_n,Y_n)}(s, t) - \Phi_{(X,y)}(s, t)| = |\mathbb{E}(e^{isX_n}(e^{itY_n} - e^{ity})) + e^{ity}\mathbb{E}(e^{isX_n} - e^{isX})|$$

On applique l'inégalité triangulaire

$$|\Phi_{(X_n,Y_n)}(s, t) - \Phi_{(X,y)}(s, t)| \leq \mathbb{E}(|e^{itY_n} - e^{ity}|) + |\mathbb{E}(e^{isX_n} - e^{isX})|$$

On a immédiatement par le théorème de Levy que le second terme à droite tend vers 0 car $X_n \to^{\mathcal{L}} X$.
Par ailleurs, $Y_n \to^{\mathbb{P}} y$ donc $Y_n \to^{\mathcal{L}} y$ et la fonction $x \mapsto |e^{itx} - e^{ity}|$ est continue bornée donc, par définition de la convergence en loi, le premier terme de droite tend également vers 0.
□

Le lemme de Slutsky autorise certaines opérations sur les limites en loi. Par exemple, $X_n \to \mathcal{N}(0, \sigma^2)$ et $\hat{\sigma}_n \to^{\mathbb{P}} \sigma$ implique

$$\frac{X_n}{\sigma} \to \mathcal{N}(0, 1)$$

ce qui sera assez utile pour les intervalles de confiance.

Ce lemme est souvent utilisé pour montrer la convergence en distribution des estimateurs. Par exemple, si un estimateur $\hat{\theta}_n \to^{\mathbb{P}} \theta$ un paramètre et une autre séquence $Z_n \to^{\mathcal{L}} Z$ une variable aléatoire, alors des combinaisons de $\hat{\theta}_n$ et Z_n peuvent être analysées en utilisant le lemme de Slutsky.

Le lemme de Slutsky implique plus explicitement que

$$X_n + Y_n \to^{\mathcal{L}} X + y$$

$$X_n Y_n \to^{\mathcal{L}} Xy$$

$$\frac{X_n}{Y_n} \to^{\mathcal{L}} \frac{X}{c},\ c \neq 0$$

En effet en reprenant la fin de la démonstration, il est facile de voir qu'en prenant $\phi : \mathbb{R} \to \mathbb{R}$ continue bornée, alors la fonction $\varphi : \mathbb{R}^2 \to \mathbb{R}$ définie par

$$\varphi(x, y) = \phi(x + y)$$

est continue bornée et donc par Slutsky

$$\mathbb{E}(\phi(X_n + Y_n)) = \mathbb{E}(\varphi(X, Y_n)) \to \mathbb{E}(\varphi(X, y)) = \mathbb{E}(\phi(X + y))$$

Le raisonnement est le même pour les autres.

Une autre conséquence du théorème de Slutsky est la Δ-méthode, permettant de transférer la propriété de normalité asymptotique via fonctionnelle différentiable.

Théorème 6.1.2 (Δ-méthode)**.** *Soit (X_n) une suite de variables aléatoires, et (r_n) suite de réels positifs tendant vers $+\infty$ tels que*

$$r_n(X_n - x) \to^{\mathcal{L}} X$$

pour un $x \in \mathbb{R}$ et X une variable aléatoire sur $\mathbb{R}$. Soit $g : \mathbb{R} \to \mathbb{R}$ une fonction différentiable en x, alors

$$r_n(g(X_n) - g(x)) \to^{\mathcal{L}} g'(x)X$$

Démonstration : Comme $r_n \to +\infty$, une première application du Lemme de Slutsky à $(r_n^{-1}, r_n(X_n - x))$ permet de montrer que $X_n \to^{\mathbb{P}} x$. On peut alors en déduire de la différentiabilité de g en x que

$$\frac{g(X_n) - g(x)}{X_n - x} \to^{\mathbb{P}} g'x$$

Le lemme de Slutsky garantit donc que

$$\left(r_n(X_n - x), \frac{g(X_n) - g(x)}{X_n - x}\right) \to^{\mathcal{L}} (X, g'(x))$$

Par continuité du produit,

$$r_n(X_n - x) \times \frac{g(X_n) - g(x)}{X_n - x} = r_n(g(X_n) - g(x)) \to^{\mathcal{L}} g'(x)X$$

□

6.1.3 Intervalle de confiance

A partir d'un estimateur T de $q(\theta)$, le but est de quantifier l'incertitude liée à cette estimation. Plus précisement, on va bâtir à partir de T des régions de $\mathbb{R}^k$ dans lesquelles le vrai paramètre $q(\theta)$ devrait se trouver avec forte probabilité.
Dans la suite, on va se placer dans un cas simple et supposer que $\Theta \subseteq \mathbb{R}$ et $q(\theta) = \theta$.

Intervalle de confiance non-asymptotique :

Définition 6.1.7 (intervalle de niveau de confiance $1-\alpha$)**.** Soit $(\mathcal{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ un modèle et $\alpha \in]0,1[$. Un intervalle de confiance (par défaut non-asymptotique) de niveau $1-\alpha$ pour θ est un couple de statistiques (T^-, T^+) tel que

$$\forall \theta \in \Theta,\ \mathbb{P}_\theta(T^- \leq \theta \leq T^+) \geq 1-\alpha$$

Lorsqu'il y a égalité, on parle d'intervalle de confiance exact.

On remarque que le cas trivial qui consisterait à prendre $T^- = -\infty$ et $T^+ = +\infty$ garanti toujours un niveau $1-\alpha$, le but implicit est donc toujours de trouver les intervalles de confiance les plus petits possibles.

On note que dans le cas où Θ n'est pas un sous-ensemble de $\mathbb{R}$, on peut définir plus généralement des régions de confiance comme des sous-ensembles aléatoires de Θ, ce qui nécessite de munir Θ d'une tribu et de vérifier certaines hypothèses de mesurabilité.

Dans le cas d'intervalle de confiance non-asymptotique, on donne 2 recettes qui correspondent à des méthodes générales.

Méthode 1 : Quantité pivotale (idéale)

Une quantité pivotale est une statistique $U(X_{1:n}, \theta)$ dont la loi sous $\mathbb{P}_\theta$ ne dépend pas de θ. Typiquement du type

$$U = \frac{T-\theta}{a}$$

avec a connu ou calculable, et U de loi connue.
Alors si $q_{\alpha/2}$ et $q_{1-\alpha/2}$ sont des quantiles de la loi de U,

$$\mathbb{P}_\theta\big(q_{\alpha/2} \leq U \leq q_{1-\alpha/2}\big) = 1-\alpha$$

se réécrit en intervalle de confiance pour θ.

Exemple 6.6. On se place dans le modèle $\mathcal{U}(]0,\theta[)^{\otimes n}$ avec

$$T_2(X_{1:n}) = \max_{i=1,\dots,n} X_i$$

Alors par invariance d'échelle, comme si $X \sim \mathcal{U}(]0,\theta[)$ alors $X/\theta \sim \mathcal{U}(]0,1[)$, pour l'échantillon i.i.d. considéré la loi de T_2/θ ne dépend pas de θ. Enfait ce phénomène est assez général : si un modèle est une famille de lois stables par un groupe de transformations (ici des dilatations) alors certaines statistiques "normalisée" deviennent des pivots. C'est précisement ce qui rend possible un intervalle

de confiance exact.
Pour $t \in (0,1)$,

$$\mathbb{P}_\theta\left(\frac{T_2}{\theta} \leq t\right) = t^n \Rightarrow \mathbb{P}_\theta\left(t \leq \frac{T_2}{\theta} \leq 1\right) = 1 - t^n$$

En prenant $t_\alpha = \alpha^{1/n}$, on obtient

$$\mathbb{P}_\theta\left(\alpha^{1/n} \leq \frac{T_2}{\theta} \leq 1\right) = 1 - \alpha$$

Si bien que

$$\left[T_2, \frac{T_2}{\alpha^{1/n}}\right]$$

est un intervalle de confiance exact de niveau $1 - \alpha$.

Dans l'exemple ci-dessus, on remarque une asymétrie naturelle.

Cette situation est assez idéale, c'est bien de savoir la reconnaitre. Un pivot apparait typiquement si

- On peut standardiser l'estimateur par une quantité connue (variance connue) ;
- On peut construire un ratio où le paramètre s'élimine (variance inconnue + Cochran) ;
- On peut utiliser une invariance du modèle (par exemple, loi uniforme$(0, \theta)$ et homothétie).

Exemple 6.7 (Moyenne d'une loi normale, variance connue). On prend

$$X_1, ..., X_n \sim^{i.i.d.} \mathcal{N}(\mu, \sigma^2)$$

avec σ^2 connue et $\mu \in \mathbb{R}$.
Alors un estimateur naturel de l'espérance est

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

On sait que

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

On va centrer-réduire pour ramener à une loi standard connue

$$Z = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \sim \mathcal{N}(0,1)$$

On prend Z comme pivot.
Soit $z_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi normale centrée réduite. Alors

$$\mathbb{P}_\mu\big(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}\big) = 1 - \alpha$$

Ce qui donne

$$\mathbb{P}_\mu\left(\overline{X}_n - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \leq \mu \leq \overline{X}_n + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\right) = 1 - \alpha$$

Donc un intervalle de confiance exact (pour tout n, ce n'est pas asymptotique) est

$$\mu \in \left[\overline{X}_n \pm \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\right]$$

Exemple 6.8 (Moyenne d'une loi normale, variance inconnue). On prend

$$X_1, ..., X_n \sim^{i.i.d.} \mathcal{N}(\mu, \sigma^2)$$

avec $\sigma^2 > 0$ inconnue et $\mu \in \mathbb{R}$.
On introduit l'estimateur de la variance empirique

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

Le résultat structurel ici est le théorème de Cochran, qui dit que dans le modèle normal

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \sim \mathcal{N}(0,1) \Rightarrow \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$$

et surtout, le coeur du théorème, ces deux quantités sont indépendantes.
On considère la quantité suivante :

$$T = \frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n}$$

et on peut réécrire T comme suit

$$T = \frac{\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}}{\sqrt{\frac{S_n^2}{\sigma^2}}} = \frac{Z}{\sqrt{\frac{U}{n-1}}}$$

où,

$$Z \sim \mathcal{N}(0,1),\ U \sim \chi^2(n-1),\ Z \perp\!\!\!\perp U$$

Par définition, celà implique donc que $T \sim t_{n-1}$ (loi de Student à $n-1$ degré de liberté). Cette loi ne dépend pas de (μ, σ^2). Si bien que T est un pivot, dit de Student.
Soit $t_{n-1,1-\alpha/2}$ le quantile $1-\alpha/2$ de t_{n-1}. Alors

$$\mathbb{P}_{\mu,\sigma^2}\big(-|t_{n-1,1-\alpha/2}| \leq T \leq |t_{n-1,1-\alpha/2}|\big) = 1-\alpha$$

En réarrangeant et en terme d'intervalle de confiance, on obtient exactement

$$\mu \in \left[\overline{X}_n \pm \frac{S_n}{\sqrt{n}} t_{n-1,1-\alpha/2}\right]$$

Cet exemple montre une chose concernant la méthodologie : quand un paramètre de nuisance (ici σ^2) gène, on cherche un pivot qui l'élimine en le remplaçant par une quantité aléatoire dont la dépendance s'annule (ici via le ratio).
C'est le prototype de beaucoup de tests/IC "classiques" (t-test).

Méthode 2 : Inégalité de concentration

Quand on ne connait pas de pivot exploitable, on majore la queue de $T-\theta$

$$\mathbb{P}_\theta\big(|T-\theta| \geq t\big) \leq \delta(t)$$

et on choisit $t = t(\alpha)$ tel que $\delta(t(\alpha)) \leq \alpha$. On obtient alors

$$\mathbb{P}_\theta\big(\theta \in [T - t(\alpha), T + t(\alpha)]\big) \geq 1 - \alpha$$

Avant de passer à la méthode, on fait quand même quelques rappels sur les inégalités de concentration basiques à savoir.

Théorème 6.1.3 (Markov et Bienaymé-Tchebychev). *Soit X une variable aléatoire réelle*

1. ***Inégalité de Markov :*** *Si $\mathbb{E}(X) < \infty$, alors pour tout $t \in \mathbb{R}$*

$$t\mathbb{P}(X \geq t) \leq \mathbb{E}(X\mathbb{1}_{X \leq t}) \leq \mathbb{E}(|X|)$$

2. ***Inégalité de Bienaymé-Tchebychev :*** *Si $\mathbb{E}(X^2) < \infty$, alors pour tout $t > 0$*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \frac{Var(X)}{t^2}$$

Pour les démonstrations de ces inégalités, on renvoie au premier chapitre de ce pdf.

Dans le cas particulier où X est une somme de variables aléatoires bornées (ce qui arrive souvent dans un cadre statistique), l'inégalité de Hoeffding est un outil très utile. On va la démontrer proprement et cela passe par le lemme de Hoeffding.

Lemme 6.1.1 (de Hoeffding). *Si X est une variable aléatoire centrée ($\mathbb{E}(X) = 0$) prenant ses valeurs dans $[a, b]$, alors*

$$\forall \lambda \geq 0,\ \psi_X(\lambda) = \log\big(\mathbb{E}\big(e^{\lambda X}\big)\big) \leq \frac{\lambda^2 (b-a)^2}{8}$$

Démonstration : Notons $\mathbb{P}_X = \mathbb{P}$, on peut réécrire

$$\psi_X(\lambda) = \log\left(\int e^{\lambda x}\mathbb{P}(dx)\right)$$

alors comme X prend ses valeurs dans $[a, b]$, il vient que $e^{\lambda x} \leq e^{\lambda b}$ $\mathbb{P}$-p.s., on peut intervertir dérivation et intégration (par convergence dominée car X est bornée), ce qui donne

$$\psi_X^{'}(\lambda) = \frac{\int x e^{\lambda x}\mathbb{P}(dx)}{\int e^{\lambda x}\mathbb{P}(dx)}$$

$$\psi_X^{''}(\lambda) = \frac{\big(\int x^2 e^{\lambda x}\mathbb{P}(dx)\big)\big(\int e^{\lambda x}\mathbb{P}(dx)\big) - \big(\int x e^{\lambda x}\mathbb{P}(dx)\big)^2}{\big(\int e^{\lambda x}\mathbb{P}(dx)\big)^2}$$

A ce moment de la démonstration, l'astuce est d'introduire une mesure Q_λ définie par :

$$Q_\lambda(dx) = \frac{e^{\lambda x}}{\int e^{\lambda x}\mathbb{P}(dx)}\mathbb{P}(dx)$$

C'est une mesure de probabilité (desnité positive et intégrale égale à 1).
Soit maintenant Y une variable aléatoire de loi Q_λ. On sait que par définition, pour toute fonction mesurable φ intégrable

$$\mathbb{E}_{Q_\lambda}(\varphi(X)) = \int \varphi(x)Q_\lambda(dx) = \frac{\int \varphi(x)e^{\lambda x}\mathbb{P}(dx)}{\int e^{\lambda x}\mathbb{P}(dx)}$$

Ainsi, en prenant $\varphi(x) = x$, on a directement que

$$\psi_X^{'}(\lambda) = \mathbb{E}_{Q_\lambda}(Y)$$

Maintenant, on voit que,

$$\psi_X^{''}(\lambda) = \frac{\big(\int x^2 e^{\lambda x}\mathbb{P}(dx)\big)\big(\int e^{\lambda x}\mathbb{P}(dx)\big) - \big(\int xe^{\lambda x}\mathbb{P}(dx)\big)^2}{\big(\int e^{\lambda x}\mathbb{P}(dx)\big)^2}$$

$$= \frac{\int x^2 e^{\lambda x}\mathbb{P}(dx)}{\int e^{\lambda x}\mathbb{P}(dx)} - \left(\frac{\int xe^{\lambda x}\mathbb{P}(dx)}{\int e^{\lambda x}\mathbb{P}(dx)}\right)^2 = \frac{\int x^2 e^{\lambda x}\mathbb{P}(dx)}{\int e^{\lambda x}\mathbb{P}(dx)} - \mathbb{E}_{Q_\lambda}(Y)^2$$

en prenant $\varphi(x) = x^2$, on obtient

$$\frac{\int x^2 e^{\lambda x}\mathbb{P}(dx)}{\int e^{\lambda x}\mathbb{P}(dx)} = \mathbb{E}_{Q_\lambda}(Y^2)$$

Par la formule de Koenig-Hyugens, il vient alors que

$$\psi_X^{''}(\lambda) = \mathrm{Var}_{Q_\lambda}(Y)$$

Maintenant que l'on a ça, on va brutalement majorer la variance de Y. On sait que $X \in [a,b]$ $\mathbb{P}$-p.s., ce qui implique de $Y \in [a,b]$ Q_λ-p.s. (Q_λ a le même support).
Si bien qu'on peut majorer :

$$\mathrm{Var}_{Q_\lambda}(Y) = \mathbb{E}_{Q_\lambda}\big((Y - \mathbb{E}(Y)^2)\big)$$

Maintenant, la moyenne minimise l'erreur quadratique, donc

$$\mathbb{E}_{Q_\lambda}\big((Y - \mathbb{E}(Y)^2)\big) \leq \mathbb{E}_{Q_\lambda}\left(\left(Y - \frac{a+b}{2}\right)^2\right)$$

pour faire plus rigoureux, on pourrait considérer la fonction $\phi(c) = \mathbb{E}\big((Y-c)^2\big)$ pour $c \in \mathbb{R}$, developper via identité remarquable et linéarité de l'espérance. On obtient alors un polynôme de degré 2 en c et donc convexe. La dérivée s'annule en $c = \mathbb{E}(Y)$ et est donc minimale en ce point. L'inégalité vient alors avec $c = \frac{a+b}{2}$.

Maintenant, soit $y \in [a,b]$, on pose le milieu $m = \frac{a+b}{2}$. Le maximum de $|y-m|$ sur $[a,b]$ est atteint aux bords a ou b. Elle vaut $\frac{b-a}{2}$. Si bien que

$$|y-m| \leq \frac{b-a}{2}$$

Ainsi, presque sûrement

$$\left(Y - \frac{a+b}{2}\right)^2 \leq \left(\frac{b-a}{2}\right)^2 = \frac{(b-a)^2}{4}$$

En prenant l'espérance et en combinant on obtient

$$\mathrm{Var}_{Q_\lambda}(Y) \leq \frac{(b-a)^2}{4}$$

Ainsi, une variable aléatoire bornée dans un intervalle de longueur $b-a$ ne peut pas avoir une variance plus grande que $\frac{(b-a)^2}{4}$. C'est un fait assez général. D'après ce fait, on a alors

$$\psi_X^{''}(\lambda) \leq \frac{(b-a)^2}{4},\ \forall \lambda \geq 0$$

Par ailleurs, on remarque que $Q_0 = \mathbb{P}$ et comme X est centrée, $\psi_X^{'}(0) = \mathbb{E}(X) = 0$.

$$\psi_X^{'}(\lambda) = \psi_X^{'}(0) + \int_0^\lambda \psi_X^{''}(t)dt \leq \int_0^\lambda \frac{(b-a)^2}{4}dt = \lambda\frac{(b-a)^2}{4}$$

Puis, on voit que $\psi_X(0) = \log \mathbb{E}(e^{0X}) = \log(1) = 0$. Si bien que de la même manière

$$\psi_X(\lambda) = \psi_X(0) + \int_0^\lambda \psi_X^{'}(t)dt \leq \int_0^\lambda t\frac{(b-a)^2}{4}dt = t^2\frac{(b-a)^2}{8}$$

□

Ici le seul saut conceptuel est l'introduction de Q_λ. C'est un changement de mesure exponentielle (tilting) qui tranforme des rapports d'intégrales en espérance ou variance, ce qui rend le contrôle immédiat ia $Y \in [a,b]$. Q_λ est une loi de rééchantillonnage biaisée : elle donne plus de poids aux grandes valeurs x si $\lambda > 0$ (car le poids est multiplié par un facteur exponentielle). On verra que c'est "naturel" d'utiliser cette transformation car on va manipuler des transformées de Laplace et des bornes de Chernoff.

Théorème 6.1.4 (Inégalité d'Hoeffding). *Soient $X_1, ..., X_n$ des variables aléatoires indépendantes telles que, pour tout $i \in \{1, ..., n\}$, $a_i \leq X_i \leq b_i$ p.s., pour $a_i, b_i \in \mathbb{R}$. En notant $S = \sum_{i=1}^n X_i$, on a*

$$\forall t \geq 0,\ \mathbb{P}\big(S - \mathbb{E}(S) \geq t\big) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

Démonstration : Quitte à translater, on peut supposer les X_i centrées avec prenant

$$\tilde{X}_i = X_i - \mathbb{E}(X_i)$$

Dans ce cas, $\mathbb{E}(\tilde{X}_i) = 0$. De plus

$$\tilde{S} = \sum_{i=1}^{n} \tilde{X}_i = S - \mathbb{E}(S)$$

Et $\tilde{X}_i$ est toujours bornée

$$a_i - \mathbb{E}(X_i) \leq \tilde{X}_i \leq b_i - \mathbb{E}(X_i)$$

L'amplitude du support ne change pas

$$(b_i - \mathbb{E}(X_i)) - (a_i - \mathbb{E}(X_i)) = b_i - a_i$$

Donc on peut se ramener à $\mathbb{E}(X_i) = 0$ et S telle que défini.
Fixons maintenant $\lambda \geq 0$. Comme $x \mapsto e^{\lambda x}$ est croissante, on a

$$\{S \geq t\} = \{e^{\lambda S} \geq e^{\lambda t}\}$$

Maintenant, on applique l'inégalité de Markov à la variable aléatoire positive $e^{\lambda S}$

$$\mathbb{P}(e^{\lambda S} \geq e^{\lambda t}) \leq \frac{\mathbb{E}\big(e^{\lambda S}\big)}{e^{\lambda t}}$$

$$e^{\lambda t}\mathbb{P}(e^{\lambda S} \geq e^{\lambda t}) \leq \mathbb{E}\big(e^{\lambda S}\big)$$

Et donc

$$e^{\lambda t}\mathbb{P}(S \geq t) \leq \mathbb{E}\big(e^{\lambda S}\big)$$

Maintenant comme S est une somme de variables aléatoires indépendantes

$$\mathbb{E}\big(e^{\lambda S}\big) = \mathbb{E}\bigg(\prod_{i=1}^{n} e^{\lambda X_i}\bigg) = \prod_{i=1}^{n} \mathbb{E}\big(e^{\lambda X_i}\big)$$

Ainsi,

$$e^{\lambda t}\mathbb{P}(S \geq t) \leq \prod_{i=1}^{n} \mathbb{E}\big(e^{\lambda X_i}\big)$$

Il reste alors à contrôler les transformées de Laplace $\mathbb{E}\big(e^{\lambda X_i}\big)$, ce que l'on fait généralement en considérant

$$\psi_{X_i}(\lambda) = \log \mathbb{E}\big(e^{\lambda X_i}\big)$$

qui est plus simple à manipuler. C'est l'objet du lemme d'Hoeffding vu plus haut.

$$\prod_{i=1}^{n} \mathbb{E}\big(e^{\lambda X_i}\big) = \exp\bigg(\sum_{i=1}^{n} \psi_{X_i}(\lambda)\bigg)$$

Donc,

$$\mathbb{P}(S \geq t) \leq \exp\Big(-\lambda t + \sum_{i=1}^{n} \psi_{X_i}(\lambda) \Big)$$

En utilisant le lemme dans la méthode de Chernoff, on obtient :

$$\psi_{X_i}(\lambda) \leq \frac{\lambda^2 (b_i - a_i)^2}{8}$$

D'où en sommant

$$\sum_{i=1}^{n} \psi_{X_i}(\lambda) \leq \frac{\lambda^2}{8} \sum_{i=1}^{n} (b_i - a_i)^2$$

Donc,

$$\forall \lambda \geq 0,\ \mathbb{P}(S \geq t) \leq \exp\Big(-\lambda t + \frac{\lambda^2}{8} \sum_{i=1}^{n} (b_i - a_i)^2 \Big)$$

Il ne reste plus qu'à optimiser la borne en λ. On pose $V = \sum_{i=1}^{n} (b_i - a_i)^2$.
On veut minimiser la quadratique

$$\phi(\lambda) = -\lambda t + \frac{\lambda^2}{8} V$$

C'est une parabole convexe (quadratique), son minimum est obtenu lorsque la dérivée s'annule. Ce fait s'accomplit en

$$\lambda^* = \frac{4t}{V} \geq 0$$

Ce qui donne,

$$\phi(\lambda^*) = -\frac{2t^2}{V}$$

On remplace :

$$\mathbb{P}(S \geq t) \leq \exp\Big(-\frac{2t^2}{V} \Big) = \exp\Big(-\frac{2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \Big)$$

Ainsi, on l'a prouvé pour des variables centrées et donc en revenant à $\tilde{S}$, obtient bien

$$\mathbb{P}(S - \mathbb{E}(S) \geq t) \leq \exp\Big(-\frac{2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \Big)$$

□

Pour information, dans la démonstration, on parle de borne de Chernoff. Ce terme signifie simplement la méthode de bornage de probabilité par transformée exponentielle. Cela combine la monotonie de l'exponentielle, l'inégalité de Markov appliquée à une variable positive puis un choix (optimisation) du paramètre λ. C'est utile car on voit que si on sait contrôler la transformée de Laplace $\mathbb{E}(e^{\lambda Z})$ alors on obtient une borne exponentielle sur la queue $\mathbb{P}(Z \geq t)$. Ensuite il ne reste qu'à choisir λ pour rendre la borne la plus petite possible

$$\mathbb{P}(Z \geq t) \leq \inf_{\lambda \geq 0} \exp\big(-\lambda t + \log \mathbb{E}(e^{\lambda t}) \big)$$

C'est exactement ce qu'il se passe dans la preuve d'Hoeffding. On contrôle $\log \mathbb{E}(e^{\lambda Z})$ via le lemme d'Hoeffding puis on optimise en λ.

Souvent on veut une inégalité bilatérale sur $|S - \mathbb{E}(S)|$, ce qui ne demande pas beaucoup plus de travail.

Théorème 6.1.5 (Inégalité de Hoeffding bilatérale)**.** *Soient $X_1, ..., X_n$ des variables aléatoires indépendantes telles que, pour tout $i \in \{1, ..., n\}$, $a_i \leq X_i \leq b_i$ p.s., pour $a_i, b_i \in \mathbb{R}$. En notant $S = \sum_{i=1}^n X_i$, on a*

$$\forall t \geq 0,\ \mathbb{P}\big(|S - \mathbb{E}(S)| \geq t\big) \leq 2 \exp\left(- \frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2} \right)$$

Démonstration : Si on veut $\mathbb{P}(|S - \mathbb{E}(S)| \geq t)$ alors on applique la borne à $S - \mathbb{E}(S)$ et $-(S - \mathbb{E}(S))$ qui sont toutes les deux sommes de variables bornées avec la même amplitude. On applique ensuite l'union bound.□

Dernière chose, cette inégalité peut se généraliser aux martingales à accroissements bornés. On l'appelle alors inégalité d'Azuma-Hoeffding. On ne la détaillera pas ici.

Nous avons maintenant la théorie nécessaire. Biensûr d'autres inégalité de concentration existe et sont utiles en statistiques mais on ne les détaillera pas ici. On va plutôt passer à des exemples pour illustrer pratiquement nos propos.

Exemple 6.9 (Taux d'éclosion des oeufs de pingouins). On collecte n oeufs de pingouins fécondés. On note

$$X_i = \mathbb{1}_{\{\text{l'oeuf } i \text{ éclôt}\}} \in \{0, 1\}$$

On modélise le vecteur aléatoire $(X_1, ..., X_n)$ comme i.i.d. de loi $\mathcal{B}(\theta)$

$$(X_1, ..., X_n) \sim \mathcal{B}(\theta)^{\otimes n},\ \theta \in (0, 1)$$

En outre, le modèle est le suivant : $\big(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), \mathcal{B}(\theta)^{\otimes n}_{\theta \in (0,1)}\big)$.
Le paramètre d'intérêt est alors θ, taux d'éclosion "théorique" de ces oeufs.
Un estimateur naturel et sans biais de θ est donné par

$$T(X_{1:n}) = \overline{X}_n$$

Sous $\mathbb{P}_\theta$, on a

$$\sum_{i=1}^n X_i \sim \mathcal{B}(n, \theta) \Rightarrow \overline{X}_n \sim \frac{1}{n} \mathcal{B}(n, \theta)$$

Ainsi, on a

$$\mathbb{E}_\theta(\overline{X}_n) = \theta \text{ et } \mathrm{Var}_\theta(\overline{X}_n) = \frac{\theta(1-\theta)}{n} \leq \frac{1}{4n}$$

Soit $\alpha > 0$, on cherche un intervalle de confiance non-asymptotique de niveau $1 - \alpha$ pour θ. Par l'inégalité de Bienaymé-Tchebychev, on a

$$\mathbb{P}_\theta(|T - \mathbb{E}_\theta T| \geq t) = \mathbb{P}_\theta(|T - \theta| \geq t) \leq \frac{\mathrm{Var}_\theta(T)}{t^2} = \frac{\theta(1-\theta)}{nt^2} \leq \frac{1}{4nt^2}$$

Si bien que,

$$\mathbb{P}_\theta(|T-\theta|<t) \geq 1-\frac{1}{4nt^2}$$

Pour obtenir un intervalle de confiance de niveau $\geq 1-\alpha$, on impose

$$\frac{1}{4nt^2} \leq \alpha \Rightarrow t \geq \frac{1}{2\sqrt{n\alpha}}$$

On pose alors

$$t_\alpha^{BT} = \frac{1}{2\sqrt{n\alpha}}$$

On a ainsi l'intervalle de confiance de niveau $\geq 1-\alpha$ suivant

$$\left[T \pm t_\alpha^{BT}\right]$$

On peut aussi utiliser l'inégalité de Hoeffding bilatérale. En effet, on sait que $X_i \in [0,1]$ presque sûrement, donc Hoeffding s'applique avec $a_i = 0$ et $b_i = 1$:

$$\mathbb{P}_\theta(|T-\theta| \geq t) \leq 2e^{-2nt^2}$$

On veut $\mathbb{P}_\theta(|T-\theta| \geq t) \leq \alpha$ donc on impose

$$2e^{-2nt^2} \leq \alpha \Leftrightarrow -2nt^2 \leq \log\left(\frac{\alpha}{2}\right) \Leftrightarrow t^2 \geq \frac{\log\left(\frac{2}{\alpha}\right)}{2n}$$

D'où,

$$t_\alpha^H = \sqrt{\frac{\log\left(\frac{2}{\alpha}\right)}{2n}}$$

Et donc, un intervalle de confiance de niveau $\geq 1-\alpha$ serait

$$\left[T \pm t_\alpha^H\right]$$

Maintenant, quand $\alpha \to 0$ l'intervalle de confiance issu de Hoeffding est bien meilleur car dans ce cadre $1/\sqrt{\alpha}$ diverge beaucoup plus vite que $\sqrt{\log(1/\alpha)}$.

Une comparaison plus pragmatique serait de dire que, supposons qu'on fixe un niveau de confiance à 90% (donc $\alpha = 0,1$). Supposons que l'on veuille une précision (longueur d'intervalle de 2%, ie $2t = 0,02$, ie $t = 0,01$).
Pour Tchebychev :

$$\frac{1}{\sqrt{n\alpha}} \leq 0,02 \Leftrightarrow \sqrt{n\alpha} \geq 50 \Leftrightarrow n \geq \frac{2500}{\alpha} = 25000$$

Tandis que Hoeffding donne :

$$2\sqrt{\frac{\log(2/\alpha)}{2n}} \leq 0,02 \Leftrightarrow n \geq \frac{\log(2/\alpha)}{2\times 10^{-4}} \sim 15000$$

Cet exemple vient d'une observation plus générale qui est la suivante

Exemple 6.10 (Moyenne empirique de variables aléatoires bornées - Hoeffding). On prend n variables aléatoires i.i.d. bornées presque sûrement $X_1, ..., X_n$

$$a \leq X_i \leq b_i \text{ p.s.}$$

et on note $\mu = \mathbb{E}(X_i)$. Alors on sait par l'inégalité de Hoeffding que

$$\mathbb{P}\big(|\overline{X}_n - \mu| \geq t\big) \leq 2\exp\bigg(-\frac{2nt^2}{(b-a)^2}\bigg)$$

On impose alors

$$2\exp\bigg(-\frac{2nt^2}{(b-a)^2}\bigg) \leq \alpha$$

ce qui donne

$$t_\alpha = (b-a)\sqrt{\frac{\log(2/\alpha)}{2n}}$$

Si bien qu'avec probabilité $\geq 1-\alpha$, on a

$$\mu \in \bigg[\overline{X}_n \pm (b-a)\sqrt{\frac{\log(2/\alpha)}{2n}}\bigg]$$

Le cas des pingouins s'obtient en prenant une loi de Bernouilli et $a, b = 0, 1$.

Comme on a dit, pour aller plus loin on peut simplement s'interesser à d'autres inégalités de concentration comme celles de type Bernstein qui est un exemple assez classique.

Intervalle de confiance asymptotique :

On se place toujours dans le cadre plus simple où $q(\theta) = \theta \in \mathbb{R}$.

Définition 6.1.8 (Intervalle de niveau de confiance asymptotique $1-\alpha$)**.** Dans un modèle i.i.d. $\big(\mathcal{X}^n, \mathcal{A}^{\otimes n}, (\mathbb{P}_\theta^{\otimes n})_\Theta\big)$ et pour $\alpha \in (0,1)$, un intervalle de confiance asymptotique de niveau $1-\alpha$ pour θ est un couple de statistiques (T_n^-, T_n^+) tel que

$$\forall \theta \in \Theta,\ \lim_{n\infty} \mathbb{P}_\theta^{\otimes n}\big(T_n^- \leq \theta \leq T_n^+\big) \geq 1-\alpha$$

7 Modèles Aléatoires (M1)

Sources : **[1]** Modèles Aléatoires, Jean-Christophe Breton, Université de Rennes - ENS Rennes, 2023-2024.
[2] Statistique Mathématique/Régression linéaire/Processus Markovien de sauts, Arnaud Guyader, Université de Rennes 2, 2015-2016.

7.1 Modèle linéaire gaussien

Le modèle linéaire gaussien est un modèle statistique fondamental utilisé pour décrire la relation entre une variable dépendante et une ou plusieurs variables indépendantes, en supposant que les erreurs ou les résidus suivent une distribution normale (gaussienne). Ce modèle est largement utilisé en statistique, en économétrie, en Machine learning et dans de nombreux autres domaines. Comprendre ce modèle est essentiel pour aborder des problèmes plus complexes.

7.1.1 Modèle linéaire simple gaussien

On considère une application affine inconnue $f(x) = \alpha x + \beta$. Si on dispose de $x_1, x_2 \in \mathbb{R}$, $y_1 = f(x_1)$ et $y_2 = f(x_2)$, alors

$$\begin{cases} x = \sin a \cos b \\ y = \sin a \sin b \end{cases} \iff \alpha = \frac{y_1 x_2 - y_2 x_1}{x_2 - x_1} \text{ et } \beta = \frac{y_2 - y_1}{x_2 - x_1}, \text{ si } (x_1, x_2) \neq \lambda(1,1)$$

Souvent, on observe pas directement $f(x) = y$ mais des valeurs bruitées :

$$Y = \alpha + \beta x + \sigma Z$$

En général, $Z \sim \mathcal{N}(0,1)$.

Dans la suite, on considère $x_1, ..., x_n$ entrées et $y_1, ..., y_n$ sorties.

$$Y_i = \alpha + \beta x_i + \sigma Z_i$$

Objectif : Estimer α et β.

On considère $u = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$, $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ et $Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_n \end{bmatrix} \in \mathcal{N}(0, I_n)$ (Vecteur gaussien standard).

$$Y = \alpha u + \beta x + \sigma Z$$

Soit $F = \text{Vect}(u, x)$ qu'on suppose de dimension 2. La stratégie est la suivante :

- Estimer α, β en projetant Y sur F.
- Estimer σ en projetant Y sur $F^\perp$.

Rappels de Géométrie euclidienne :

On pose $F = \text{Vect}(x - \overline{x}u, u)$, $x - \overline{x}u$ est orthogonal à u au sens euclidien et avec

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n} < x, u > = < x, \frac{u}{||u||^2} >$$

On a aussi, on notant p_F la projection sur le sous-espace F,

$$\forall z \in \mathbb{R}^n,\ z = p_F(z) + p_{F^\perp}(z)$$

Comme $\left(\frac{u}{||u||}, \frac{x - \overline{u}x}{||x - \overline{x}u||}\right)$ est une base orthonormée de F, on a

$$p_F(z) = < z, \frac{u}{||u||} > \frac{u}{||u||} + < z, \frac{x - \overline{x}u}{||x - \overline{x}u||} > \frac{x - \overline{x}u}{||x - \overline{x}u||} = \overline{z}.\frac{u}{||u||} + b(z)(x - \overline{x}u)$$

où

$$b(z) = < z, \frac{x - \overline{x}u}{||x - \overline{x}u||^2} > = \frac{\sum z_i(x_i - \overline{x})}{\sum (x_i - \overline{x})^2}$$

On note $\hat{Y} = p_F(Y) = \overline{Y}u + b(Y)(x - \overline{x}u) = (\overline{Y} - \overline{x}b(Y))u + b(Y)x$ et, on pose $A = (\overline{Y} - \overline{x}b(Y))$ et $B = b(Y)$.

$$\delta^2 = \frac{1}{n-2}\frac{i=1}{n}(x_i - \overline{x})^2 = \frac{1}{n-2}||Y - \hat{Y}||^2 = \frac{1}{n-2}||p_{F^\perp}(Y)||^2$$

Rappels probabiliste :

On a $\Gamma(1, \lambda) \sim \mathcal{E}(\lambda)$ et $\Gamma\left(\frac{d}{2}, \frac{1}{2}\right) \sim \chi^2(d)$.

Théorème 7.1.1 (de Cochran). *Soit W un vecteur gaussien centré $\mathcal{N}(0, \sigma^2 I_d)$. Soit une décomposition orthogonale :*

$$\mathbb{R}^d = V_1 \oplus ... \oplus V_k$$

Alors, $p_{V_i}(W)$ vecteur gaussien orthogonal indépendant de $p_{V_j}(W)$ et

$$\frac{||p_{V_i}(W)||^2}{\sigma^2} \sim \chi^2(dim V_i)$$

Fin des rappels. On garde aussi les notations.

Théorème 7.1.2. *Les variables aléatoires A, B et δ^2 sont orthogonaux et on a*

$$\hat{Y} \sim \mathcal{N}(\alpha + \beta\overline{x}, \frac{\sigma^2}{n})$$

$$\frac{n-2}{\sigma^2}\delta^2 \sim \chi^2(n-2)$$

$$B \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{\sum(x_i - \overline{x})^2}\right)$$
$$A \sim \mathcal{N}\left(\alpha, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum(x_i - \overline{x})^2}\right)\right)$$

Démonstration : On a $Y = \alpha u + \beta x + \sigma^2 Z$. De plus

$$\overline{Y} = < Y, \frac{u}{||u||^2} >$$

$$B = < Y, \frac{x - \overline{x}u}{||x - \overline{x}u||} >$$

Alors, $(\overline{Y}, B)$ est un vecteur gaussien.
$a\overline{Y} + bB = a < Y, \frac{u}{||u||^2} > + b < Y, \frac{x-\overline{x}u}{||x-\overline{x}u||} >$ est gaussien.
De plus, $\mathrm{Cov}(\overline{Y}, B) = 0$ car

$$\mathrm{Cov}(\overline{Y}, B) = \mathrm{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i, \sum_{j=1}^{n} \frac{Y_j(x_j - \overline{x})}{||x - \overline{x}u||^2}\right) = \frac{1}{n||x - \overline{x}u||^2} \sum_{i,j=1}^{n} \mathrm{Cov}(Y_i, Y_j)(x_j - \overline{x})$$

et, comme $\alpha x + \beta x_i$ et $\alpha u + \beta x_j$ sont déterministes

$$\mathrm{Cov}(Y_i, Y_j) = \mathrm{Cov}(\alpha u + \beta x_i + \sigma Z_i, \alpha u + \beta x_j + \sigma Z_j) = \mathrm{Cov}(\sigma Z_i, \sigma Z_j)$$
$$= \sigma^2 \mathrm{Cov}(Z_i, Z_j) =^{\perp} \sigma^2 \delta_{i,j}$$

Donc, $\overline{Y}$ et B sont indépendants car gaussien.
Ainsi, $\overline{Y} = \alpha + \beta x + \sigma\overline{Z} \sim \mathcal{N}(\alpha + \beta\overline{x}, \frac{\sigma^2}{n})$ Puis, $B = < Y, \frac{x-\overline{x}u}{||x-\overline{x}u||} >$ et,

$$\mathbb{E}(B) = \mathbb{E}\left(\sum_{i=1}^{n} \frac{Y_i(x_i - \overline{x})}{||x - \overline{x}u||^2}\right) = \sum_{i=1}^{n} \frac{\mathbb{E}(Y_i)(x_i - \overline{x})}{||x - \overline{x}u||^2} = \sum_{i=1}^{n} \frac{\alpha + \beta x_i)(x_i - \overline{x})}{||x - \overline{x}u||^2}$$

$$= \beta \sum_{i=1}^{n} \frac{x_i(x_i - \overline{x})}{||x - \overline{x}u||^2} - \beta \sum_{i=1}^{n} \frac{\overline{x}(x_i - \overline{x})}{||x - \overline{x}u||^2} = \beta \sum_{i=1}^{n} \frac{x_i(x_i - \overline{x})}{||x - \overline{x}u||^2}$$

car $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$.

$$\mathbb{E}(B) = \beta . \frac{1}{||x - \overline{x}u||^2} \sum_{i=1}^{n}(x_i - \overline{x})^2 = \beta$$

Or,

$$\mathrm{Var}(B) = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{||x - \overline{x}u||} \mathrm{Var}(Y_i) = \frac{\sigma^2}{||x - \overline{x}u||^2}$$

D'où,

$$B \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{||x - \overline{x}u||^2}\right)$$

Maintenant, $A = \overline{Y} - B\overline{x}$ de loi normale car $(B, \overline{Y})$ est gaussien.

$$\mathbb{E}(A) = \mathbb{E}(\overline{Y}) - \mathbb{E}(B)\overline{x} = \alpha + \beta\overline{x} - \beta\overline{x} = \alpha$$

Puis, comme B et $\overline{Y}$ sont indépendants

$$\mathrm{Var}(A) = \mathrm{Var}(\overline{Y}) + \overline{x}^2\mathrm{Var}(B) = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{||x - \overline{x}u||^2}\right)$$

Si bien que,

$$A \sim \mathcal{N}\left(\alpha, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{||x - \overline{x}u||^2}\right)\right)$$

Pour σ^2, on applique Cochran à $Z \sim \mathcal{N}(0, I_n)$ et F sous-espace vectoriel de dimension 2.

$$\hat{Z} = p_F(Z) \perp\!\!\!\perp p_{F^\perp}(Z) = Z - p_F(Z) = Z - \hat{Z}$$

et, $||p_{F^\perp}(Z)||^2 \sim \chi^2(n-2)$. Donc, $||Z - \hat{Z}||^2 \sim \chi^2(n-2)$.

$$Y = \alpha u + \beta x + \sigma Z$$

$$\hat{Y} = p_F(Y) = \alpha u + \beta x + \sigma\hat{Z}$$

Ainsi, $Y - \hat{Y} = \sigma(Z - \hat{Z}) \Rightarrow \dfrac{||Y - \hat{Y}||^2}{\sigma^2} = ||Z - \hat{Z}||^2$.

Et, $\dfrac{(n-2)\delta^2}{\sigma^2} = \dfrac{||Y - \hat{Y}||^2}{\sigma^2}$. Pour terminer, on observe :

$$\sigma(\delta^2) = \sigma(Y - \hat{Y}) = \sigma(Z - \hat{Z}) \perp\!\!\!\perp \sigma(\hat{Z})$$

D'où,

$$\sigma(\delta^2) \perp\!\!\!\perp \sigma(\overline{Z}, b(\overline{Z})) = \sigma(\overline{Y}, b(Y))$$

Si bien que,

$$\delta^2 \perp\!\!\!\perp Y, B$$

□

De ce théorème on tire des conséquences statistiques :

- A est un estimateur sans biais de α ;
- B est un estimateur sans biais de β ;
- δ^2 est un estimateur sans biais de σ^2. En effet

$$\mathbb{E}\left(\frac{(n-2)\delta^2}{\sigma^2}\right) = \mathbb{E}(\chi^2(n-2)) = n - 2$$

- $\overline{Y}$ est un estimateur sans biais de $\alpha + \beta\overline{x}$.

On notera une subtilité dans le fait que δ n'est pas un estimateur sans biais de σ. En effet, la fonction racine étant strictement concave, par l'inégalité de Jensens on a pas d'égalité.

De plus, on notera que A et B ne sont pas $\perp\!\!\!\perp$.

7.1.2 Modèle linéaire général

On considère la fonction linéaire

$$f(x) = \sum_{i=1}^{p} \beta_i x_i =< \beta, x >= \beta^T.x$$

On suppose que l'image d'une entrée x est observée avec un bruit réalisé par une variable aléatoire Z et on s'intéresse alors à l'observation bruitée

$$Y_i = \sum_{i=1}^{p} \beta_j x_{i,j} + \overline{Z}_i$$

En notant $Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_p \end{bmatrix}$, $X = (x_{i,j})_{i,j}$ où $x_{i,j}$ désigne l'entrée numéro i, $Z = \begin{bmatrix} \overline{Z}_1 \\ \vdots \\ \overline{Z}_p \end{bmatrix}$.

$$Y = X\beta + Z$$

L'objectif est donc de déterminer $\beta_1, ..., \beta_p$ et Z (quand c'est possible).

Hypothèse 1 : Z est entrée. Alors

$$\mathbb{E}(Y) = \mathbb{E}(X\beta) = X\beta$$

Hypothèse 2 : $\mathrm{Var}(Z_i) = \sigma^2 < +\infty$ (Homoscédastique).
Souvent on suppose Z gaussien $\sim \sigma\mathcal{N}(0, I_p)$

$$Y = X\beta + Z \sim \mathcal{N}(X\beta, \sigma^2 I_p)$$

Hypothèse 3 : $n > p$, $X \in \mathcal{M}_{n,p}$, $\mathrm{rg}(X) = p$.
C'est-à-dire que, en notant X^i les colonnes de X, $F = \mathrm{Vect}(X^1, ..., X^p)$ est de dimension p.
Dans ce cas, $X^T X$ est inversible.

Exemple 7.1. Le modèle constant : $p = 1$. Alors $X = (1)$, sans intérêt.

Exemple 7.2. Modèle simple : $p = 2$. $x_{i,1} = 1$ et $x_{i,2} = x_i$

Exemple 7.3. Modèle d'analyse à 2 variables.

Estimation de β par moindres carrés :

Sous les hypothèses sur la loi du bruit Z, on estime β en prenant :

$$\hat{\beta} = \mathrm{Argmin}||Y - X\beta||^2$$

Remarque 7.1. L'Argmin est le point en lequel le min se réalise. On veut que $X\beta$ soit proche de Y, donc on veut minimiser $||Y - X\beta||^2$.

Proposition 7.1.1. *L'unique solution est* $\hat{\beta} = (X^T X)^{-1} X^T Y$ *tel que* $Cov(Z) = \sigma^2 I_n$.

Démonstration : On a

$$||Y - X\beta||^2 = ||Y||^2 - 2 < X^T Y, \beta > + ||X\beta||^2$$

Puis,

$$\nabla_\beta ||Y - X\beta||^2 = -2 < X, Y - X\beta >= -2X^T Y + 2(X^T X)\beta = 0 \Longleftrightarrow \beta = (X^T X)^{-1} X^T Y$$

On regarde la hessienne pour s'assurer qu'il y a bien un minimum :

$$\nabla^2_\beta ||Y - X\beta||^2 = 2(X^T X)$$

qui est une matrice positive.
Donc $\hat{\beta} = (X^T X)^{-1} X^T Y$ réalise bien le minimum. □

On remarque que,

$$\mathbb{E}(\hat{\beta}) = \mathbb{E}((X^T X)^{-1} X^T Y) = (X^T X)^{-1} X^T \mathbb{E}(Y) = \beta$$

$$\text{Cov}(\hat{\beta}) = \text{Cov}((X^T X)^{-1} X^T Y) = \text{Cov}(AY) = \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}$$

Et on rappelle biensûr que

$$\text{Cov}(U) = \Sigma \Rightarrow \text{Cov}(AU) = A\Sigma A^T$$

Lorsque la structure de la covariance du bruit est de la forme $\text{Cov}(Z) = \sigma^2 \Sigma$, où Σ est une matrice carrée définie positive connue, alors on peut encore déterminer un estimateur des moindres carrés. Il est de la forme

$$\hat{\beta} = (X^T \Sigma X)^{-1} X^T \Sigma^{-1} X$$

Modèle linéaire gaussien :

On suppose que $Z \sim \mathcal{N}(0, \sigma^2 I_n)$. On a alors vu que $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$.
Pour détailler l'estimateur de β et σ^2, on projette sur $F = \text{Vect}(X^1, ..., X^p)$ et $F^\perp$ et on utilise les propriétés gaussienne.

Proposition 7.1.2 (Cas gaussien)**.**

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^T \Sigma^{-1} X)^{-1})$$

Démonstration : $\hat{\beta} = (X^T X)^{-1} X^T Y$.□

On passe sur le cas des moindres carrés généralisés.

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^T\Sigma^{-1}X)^{-1})$$

Lorsque les X^j sont $\perp$ deux à deux, alors

$$X^TX = \text{diag}(\alpha_1, ..., \alpha_p)$$

où,

$$\alpha_j = \sum_{i=1}^{n} \alpha_{i,j}^2$$

Alors,

$$\hat{\beta}_j \sim \mathcal{N}(0, \alpha_j^{-1})$$

où les $\hat{\beta}_j$ sont indépendants.

Proposition 7.1.3. *Soit $H = X(X^TX)^{-1}X^T$. Alors H est la projection orthogonale sur F.*

Démonstration : H est évidemment symétrique. $H^TH = H$ donc H est une projection. H est une projection orthogonale.
Soit $u \in \mathbb{R}^p$,

$$< Hu, Hu >=< H^THu, u >=< Hu, u >\Rightarrow< Hu, Hu - u >= 0$$

$$u = Hu + (u - Hu) = p_F(u) + p_{F^\perp}(u)$$

On montre par des arguments de dimensions que $\Im(H) = F$.

$$Hu = X(X^TX)^{-1}X^Tu = Xv = X^1v_1 + ... + X^pv_p \in F$$

On remarque aussi que $I - H$ est la projection orthogonale sur $F^\perp$. □

Définition 7.1.1. On appelle vecteur des prédictions, le vecteur

$$\hat{Y} = X\hat{\beta}$$

On appelle vecteur des résidus, le vecteur

$$Y - \hat{Y} = \hat{Z}$$

Proposition 7.1.4.

1. $\hat{Y} \sim \mathcal{N}(X\beta, \sigma^2H)$;
2. $\hat{Z} \sim \mathcal{N}(0, \sigma^2(I_n - H))$;
3. $\hat{Y} \perp\!\!\!\perp \hat{Z}$ *et* $\hat{\beta} \perp\!\!\!\perp \hat{Z}$;
4. $\frac{||\hat{Z}||^2}{\sigma^2} \sim \chi^2(n - p)$.

Démonstration : Pour le premier point, on a par définition que $\hat{Y} = X\hat{\beta}$ avec $\beta \sim \mathcal{N}(\beta, \sigma^2(X^TX)^{-1})$

$$\Rightarrow \hat{Y} \sim \mathcal{N}(X\beta, X\{\sigma^2(X^TX)^{-1}\}X^T) \sim \mathcal{N}(X\beta, \sigma^2 H)$$

Pour le second point, on a $\hat{Z} = Y - \hat{Y} = p_{F^\perp}(Y)$ où $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$.
D'où

$$\hat{Z} \sim \mathcal{N}((I_n - H)X\beta, (I_n - H)\sigma^2 I_n (I_n - H)^T)$$

et d'une part, $(I_n - H)X\beta = 0$ car $X\beta = X^1\beta_1 + ... + X^p\beta_p \in F$.
D'autre part, $(I_n - H)\sigma^2 I_n (I_n - H)^T = \sigma^2(I_n - H)$ car $I_n - H$ est symétrique et c'est une projection.
Pour le troisième point, $(\hat{Y}, \hat{Z}) = (p_F(Y), p_{F^\perp}(Y))$ (Théorème de Cochran)

$$\hat{Y} = p_F(Y) \perp\!\!\!\perp p_{F^\perp}(Y) = \hat{Z}$$

On a,

$$X^T\hat{Y} = X^TX\hat{\beta} \Rightarrow \hat{\beta} \perp\!\!\!\perp \hat{Z}$$

avec $X^T\hat{\beta} = (X^TX)^{-1}X^TY$ où $Y \perp\!\!\!\perp \hat{Z}$. $\square$

On verra le dernier point plus tard.

Conséquences statistiques :

1. $\hat{\beta}$ est un estimateur sans biais de β ;
2. $\hat{\beta}$ est un estimateur sans biais de $X\beta$;
3. $\sigma^2 = \dfrac{||\hat{Z}||^2}{n-p}$ est un estimateur sans biais de σ^2 ;

Or, $\hat{Z} \sim \mathcal{N}(0, \sigma^2(I_n - H))$, donc $\dfrac{||\hat{Z}||^2}{\sigma^2} \sim \chi^2(n-p)$ par le théorème de Cochran.
Si bien que

$$\mathbb{E}\left(\frac{(n-p)\delta^2}{\sigma^2}\right) = \mathbb{E}(\chi^2(n-p)) = n - p$$

C'est-à-dire,

$$\mathbb{E}(\delta^2) = \sigma^2$$

4. $(\hat{\beta}, \hat{Y}) \perp\!\!\!\perp (\delta^2, ||\hat{Z}||^2)$;
5.

7.1.3 Tests sur le modèle linéaire

7.2 Arbres de Galton-Watson

7.2.1 Famille de Galton-Watson

Soit X de loi quelconque sur $\mathbb{N}$ notée μ.

$$\mathbb{P}(X = k) = p_k$$

On a à l'instant initial $Z_0 = z \in \mathbb{N}$ individus qui peuvent avoir k avec probabilité

$$\mathbb{P}(k \text{ enfants}) = \mathbb{P}(X = k) \sim^{\mathcal{L}} \mu$$

Chaque individus peut avoir k enfants à la génération suivante. On note :

- $(Z_n)_n$ la suite de la population ;
- $X_{n,k}$ le nombre d'enfants de l'individu k à la génération $n-1 \to n$.

Alors, si on prend les notations précédentes avec $(X_{n,k})_{n,k}$ indépendante et identiquement distribuée de même loi que X :

$$Z_{n+1} = \sum_{k=1}^{Z_n} X_{n,k}$$

On remarque tout de suite que si pour un certain n_0, $Z_{n_0} = 0$ alors pour tout $n \geq n_0$, $Z_n = 0$.

On peut aussi commencer par remarquer plus généralement que $\mathcal{L}(Z_{n+1}|Z_0, ..., Z_n) = \mathcal{L}(Z_{n+1}|Z_n)$ et $\mathcal{L}(Z_{n+1}|Z_n = k) = \mu^{*k}$. On sous-entend $Z_{n+1} = X_1 + ... + X_n$ avec (X_i) indépendantes et identiquement distribuée de loi μ.

Proposition 7.2.1. *La suite (Z_n) est une chaîne de Markov sur $\mathbb{N}$ de noyau de transition*

$$P(k, .) = \mu^{*k}$$

et 0 est un état absorbant.
Lorsque $p_0 > 0$ tous les états sont transitoires.

Définition 7.2.1 (Temps d'extinction)**.** Le temps d'extinction de la population est

$$T = \inf\{n \geq 0 | Z_n = 0\}$$

avec $\inf \emptyset = \infty$.

Exemple 7.4. $Z_0 = 1$ et $\mu = \delta_k$, $Z_{n+1} = kZ_n$. Faire un arbre.
Ce modèle est très simple et ne prend pas en compte les différents types (par exemple, masculin ou féminin) ou des migrations (arrivée d'individus extérieurs).

Dans la suite, on se ramène à $Z_0 = 1$ car conditionnellement à $\{Z_0 = k\}$, la suite à la même loi que la somme de k copies indépendantes et identiquement distribuées de la suite partant de $Z_0 = 1$.

Proposition 7.2.2 (de branchement)**.** *Elle dit entre-autre que si un évènement se passe pour l'arbre entier (par exemple, l'arbre meurt), alors cela doit se passer pour tous les sous-arbres et vice-versa.*

$$Z^{(x+y)} \sim Z^{(x)} + Z^{(y)}$$

Proposition 7.2.3. *Si $p_0 = 0$ et $p_1 < 1$ alors*

$$\mathbb{P}(Z_n \to +\infty) = 1$$

C'est-à-dire que l'arbre explose.

Proposition 7.2.4. *Si $p_0 + p_1 = 1$ alors*

$$\mathbb{P}(Z_n \to 0) = 1$$

C'est-à-dire que presque sûrement la population s'éteint en une durée $T \sim \mathcal{G}(p)$.

Ces observations conduisent à se ramener dans la suite à la structure générique où

$$Z_0 = 1,\ 0 < p_0 \geq p_0 + p_1 < 1 \text{ et } \forall k \geq 0,\ p_k < 1$$

où le dernier $p_k < 1$ est impliqué par $0 < p_0 < 1$.

Maintenant, on note

$$m = \mathbb{E}(Z_1) = \sum_{k=0}^{+\infty} k p_k = \mathbb{E}(\mu)$$

$$\sigma^2 = \mathrm{Var}(Z_1) = \mathbb{E}(Z_1^2) - m^2$$

Proposition 7.2.5. *1. Si $m < \infty$, $\mathbb{E}(Z_n) = m^n$;*

2. Si $\sigma^2 < \infty$,

$$Var(Z_n) = \frac{\sigma^2 m^n (m^n - 1)}{m^2 - m} \text{ si } m \neq 1, \text{ et } n\sigma^2 \text{ si } m = 1$$

Démonstration : Pour le premier point, c'est du calcul

$$\mathbb{E}(Z_{n+1}) = \mathbb{E}(\mathbb{E}(Z_{n+1}|Z_n)) = \mathbb{E}\left(\sum_{k=1}^{Z_n} \mathbb{E}(X_{n,k}|Z_n) \right)$$

et comme $X_{n,k} \perp\!\!\!\perp Z_n$,

$$\mathbb{E}(Z_{n+1}) = \mathbb{E}\left(\sum_{k=1}^{Z_n} m \right) = m\mathbb{E}(Z_n)$$

On conclut par une récurrence immédiate que

$$\mathbb{E}(Z_{n+1}) = m^{n+1}$$

Pour le deuxième point, et par ce qui précède

$$\mathrm{Var}(Z_{n+1}) = \mathrm{Var}(\mathbb{E}(Z_{n+1}|Z_n)) + \mathbb{E}(\mathrm{Var}(Z_{n+1}|Z_n)) = m^2 \mathrm{Var}(Z_n) + \mathbb{E}\left(\mathrm{Var}\left(\sum_{k=1}^{Z_n} X_{n,k} \middle| Z_n \right) \right)$$

Or, $X_{n,k} \perp\!\!\!\perp Z_n$ et les $(X_{n,k})$ étant indépendants et identiquement distribués

$$\mathrm{Var}\left(\sum_{k=1}^{Z_n} X_{n,k} \middle| Z_n\right) = \sum_{k=1}^{Z_n} \mathrm{Var}(X_{n,k}|Z_n) = \sum_{k=1}^{Z_n} \sigma^2 = \sigma^2 Z_n$$

C'est-à-dire,

$$\mathrm{Var}(Z_{n+1}) = m^2 \mathrm{Var}(Z_n) + \sigma^2 \mathbb{E}(Z_n) = m^2 \mathrm{Var}(Z_n) + \sigma^2 m^n$$

□

Définition 7.2.2 (Régimes)**.** Pour une loi de reproduction μ, de moyenne $m < +\infty$, on dit que l'arbre de Galton-Watson est :

- Sous-critique, si $m < 1$ i.e. $\mathbb{E}(Z_n) = m^n \to 0$ en décroissance ;
- Critique, si $m = 1$ i.e. $\mathbb{E}(Z_n) = m^n = 1$ pour tout n ;
- Sur-critique, si $m > 1$ i.e. $\mathbb{E}(Z_n) \to +\infty$

Tout sera expliqué en détails dans les prochaines sous-parties.

On réintroduit la fonction génératrice de loi μ, qui sera très important dans cette section.

$$G : [0,1] \to [0,1]$$

avec

$$G(s) = \mathbb{E}(s^{Z_1}) = \sum_{k=0}^{+\infty} p_k s^k \quad (Z_1 \sim \mu \text{ si } Z_0 = 1)$$

On remarque que G est $\mathcal{C}^\infty$ sur $[0,1[$ et $\mathcal{C}^0$ sur $[0,1]$. Si $m < \infty$ alors G est $\mathcal{C}^1$ sur $[0,1]$. Plus généralement, si μ a un moment d'ordre d'ordre r, alors $G \in \mathcal{C}^r([0,1])$ et

$$G^{(r)}(1) = \mathbb{E}(Z_1(Z_1 - 1)...(Z_1 - r + 1))$$

En particulier,

$$G'(1) = m$$

$$G''(1) = \mathbb{E}(Z_1^2) - \mathbb{E}(Z_1) = \mathrm{Var}(Z_1) + \mathbb{E}(Z_1)^2 - E(Z_1) = \sigma^2 + m(m-1)$$

Proposition 7.2.6. *La fonction génératrice G_n de Z_n vérifie*

$$G_n = G^{\circ n}$$

Démonstration : On va raisonner par récurrence,

$$G_1 = G_{Z_1} = G_\mu = G^{\circ 1}$$

Pour passer de n à $n+1$,

$$G_{n+1}(s) = \mathbb{E}(s^{Z_{n+1}}) = \mathbb{E}(\mathbb{E}(s^{Z_{n+1}}|Z_n)) = \mathbb{E}\left(\prod_{k=1}^{Z_n} \mathbb{E}(s^{X_{n,k}}|Z_n)\right)$$

$$=^{s^{X_{n,k}} \perp\!\!\!\perp Z_n} \mathbb{E}(G(s)^{Z_n}) = G_n(G(s)) = (G_n \circ G)(s)$$

On a alors $Z_{n+1}|Z_n \sim \mu^{*n}$ donc

$$\mathbb{E}(\mathbb{E}(s^{Z_{n+1}}|Z_n)) = G^{\circ n+1}(s)$$

□

7.2.2 Probabilité d'extinction

Théorème 7.2.1. *Presque sûrement, soit la famille s'éteint soit la famille explose.*

Démonstration : Comme $p_0 \neq 0$, tout état est transitoire. Si la chaîne est bornée et passe un temps fini en tout $n \neq 0$ cela oblige à atteindre 0. Sinon elle n'est pas bornée et explose. □

On note la probabilité d'extinction

$$\rho_0 = \mathbb{P}(T < \infty) = 1 - \mathbb{P}(\lim Z_n = +\infty)$$

Proposition 7.2.7.

$$\rho_0 = \lim \mathbb{P}(Z_n = 0) = \lim G^{\circ n}(0)$$

Démonstration : On pose

$$\beta = \mathbb{P}(T < \infty)$$

On a

$$\{T < \infty\} = \bigcup_{n\geq 0} \{T \leq n\} = \bigcup_{n\geq 0} \{Z_n = 0\}$$

D'où,

$$\rho_0 = \mathbb{P}\left(\bigcup_{n\geq 0} \{Z_n = 0\} \right) = \lim \mathbb{P}(Z_n = 0)$$

Or, $G_n(0) = \mathbb{P}(Z_n = 0)$. D'où le résultat. □

Théorème 7.2.2. *On a les 2 situations suivantes*

- *Si $m \leq 1$, alors $\rho_0 = 1$;*
- *Si $m > 1$, alors ρ_0 est l'unique point fixe de G sur $]0, 1[$.*

Démonstration : D'une part

$$\mathbb{P}(T \leq n) = \mathbb{P}(Z_n = 0) = G_n(0) = G^{\circ n}(0)$$

$$\mathbb{P}(T \leq n+1) = G^{\circ n+1}(0) = G(G^{\circ n}(0)) = G(\mathbb{P}(T \leq n))$$

Maintenant, comme G est continue sur $[0, 1]$ et comme $\rho_0 = \lim \mathbb{P}(T < n)$, par passage à la limite

$$G(\rho_0) = \rho_0$$

Par suite, on remarque que $G'' \geq 0$ donc G est convexe sur $[0, 1]$. Comme $0 < p_0 + p_1 < 1$, il existe $k \geq 2$ tel que $p_k > 0$.

Donc, $G'' > 0$ i.e. G est strictement convexe.
La convexité exige que la courbe de G n'intersecte pas celle de $x \mapsto x$ plus d'une fois.
Quand $m > 1$, $G'(1) > 1$ et la pente de G en 1 étant plus grande que celle de la droite $x \mapsto x$, le graphe doit être en dessous de la droite en 1^- et G intersecte $\mathrm{id}(x) = x$ sur $[0, 1[$ par TVI.
Il reste à voir que si $m > 1$, ρ_0 est le plus petit point fixe de G, noté s. Pour cela, en montrant que

$$\forall n,\ \mathbb{P}(T \leq n) \leq s \Rightarrow \rho_0 \leq s \text{ et } \rho_0 \in \{s, 1\}$$

Puisque $Z_0 = 1$, $T > 0$. On suppose $\mathbb{P}(T \leq n) \leq s$.

$$\mathbb{P}(T \leq n+1) = G(\mathbb{P}(T \leq n)) \leq G(s) = s$$

D'où le résultat. □

On introduit maintenant

$$Y_n = \frac{Z_n}{\mathbb{E}(Z_n)} = \frac{Z_n}{m^n},\ n \geq 0$$

On pose

$$\mathcal{F}_n = \sigma(Y_0, ..., Y_n) = \sigma(Z_0, ..., Z_n),\ n \geq 1 \Rightarrow (\mathcal{F}_n)_{n \geq 0} \text{ est la filtration canonique de } (Y_n)_n$$

Proposition 7.2.8. *$Y = (Y_n)_n$ est une martingale positive, donc elle converge presque sûrement.*

Démonstration : On a la filtration canonique donc adaptée

$$\mathbb{E}(Y_{n+1}|\mathcal{F}_n) = \mathbb{E}\left(\frac{Z_{n+1}}{m^{n+1}}\middle| Z_0, ..., Z_n\right) = \frac{1}{m^{n+1}}\mathbb{E}(Z_{n+1}|Z_n) = \frac{Z_n}{m^n} = Y_n$$

$Y \geq 0$: OK
Par propriété sur les martingales positives, elles convergent presque sûrement. □

Corollaire 7.2.1. *Presque sûrement sur $\{Y_\infty\}$,*

$$Z_n =_{n \to +\infty} \mathcal{O}(m^n)$$

7.2.3 Cas sous-critique

Théorème 7.2.3. *Dans le cas d'un arbre de Galton-Watson sous-critique, $\mathcal{L}(Z_n|Z_n > 0)$ converge vers une loi ν de la fonction génératrice H qui vérifie*

$$(H-1)(G(s)) = m(H(s)-1)$$

Démonstration : On note $H_n = G_{\nu_n}$ avec $\nu_n \sim \mathcal{L}(Z_n|Z_n > 0)$. D'abord, $H_n(1) = 1 \to H(1) = 1$.
Ensuite, pour $s \in [0, 1[$,

$$1 - H_n(s) = 1 - \mathbb{E}(s^{Z_n}|Z_n > 0) = 1 - \frac{\mathbb{E}(s^{Z_n}\mathbb{1}_{Z_n > 0})}{\mathbb{P}(Z_n > 0)} = 1 - \frac{\mathbb{E}(s^{Z_n}) - \mathbb{E}(s^{Z_n}\mathbb{1}_{Z_n = 0})}{1 - \mathbb{P}(Z_n = 0)}$$

$$= 1 - \frac{\mathbb{E}(s^{Z_n}) - \mathbb{P}(Z_n = 0)}{1 - \mathbb{P}(Z_n = 0)} = \frac{\mathbb{P}(Z_n > 0) + \mathbb{P}(Z_n = 0) - G_n(s)}{1 - \mathbb{P}(Z_n = 0)} = \frac{1 - G_n(s)}{1 - \mathbb{P}(Z_n = 0)}$$

Maintenant, $G_{n+1} = G \circ G_n$ et 1 est le seul point fixe de G (car $m < 1$), on a $G_n(s) \to 1$ pour tout s.

On a G_n croissante, $\frac{1 - G(s)}{1 - s}$croit car G est convexe.

$$\frac{1 - H_{n+1}(s)}{1 - H_n(s)} = \frac{1 - G_{n+1}(s)}{1 - G_{n+1}(0)} \times \frac{1 - G_n(0)}{1 - G_n(s)} = \frac{1 - G(G_n(s))}{1 - G_n(s)} \times \frac{1 - G_n(0)}{1 - G(G_n(0))}$$

$$= \frac{r(G_n(s))}{r(G_n(0))}$$

Ainsi, $(1 - H_n(s))_n$ est une suite croissante et majorée donc converge.

$$1 - H_n(G(s)) = \frac{1 - G_n(G(s))}{1 - G_n(0)} = \frac{1 - G_{n+1}(0)}{1 - G_n(0)} \times \frac{1 - G_{n+1}(s)}{1 - G_{n+1}(0)}$$

$$= \frac{G(1) - G(G_n(0))}{1 - G_n(0)} \times \frac{1 - G_{n+1}(s)}{1 - G_{n+1}(0)} \to G'(1)(1 - H(s)) = m(1 - H(s))$$

La convergence en loi vient de la convergence en fonction génératrice.□

On a aussi dans le cas sous-critique

Proposition 7.2.9. $\forall i, n \in \mathbb{N}^*$,

$$\mathbb{P}(Z_n > 0 | Z_0 = i) \geq \frac{i(1-m)m^{n+1}}{\sigma^2(1-m^n) + m^{n+1}(1-m)} \left(1 - \frac{(i-1)m^n}{2}\right)$$

Et,

$$\mathbb{P}(Z_n > 0 | Z_0 = i) \leq im^n$$

$$\mathbb{P}(Z_n > 0 | Z_0 = i) = \mathbb{E}(\mathbb{1}_{Z_n > 0} | Z_0 = i) \leq \mathbb{E}(Z_n | Z_0 = i)$$

Ainsi,

$$i\frac{m(1-m)m^n}{\sigma^2} \leq \mathbb{P}(Z_n > 0 | Z_0 = i) \leq im^n$$

Proposition 7.2.10. *On a*

$$\mathbb{E}(T | Z_0 = i) \sim_{i,m \to +\infty} \frac{\ln(i)}{\ln(m)}$$

7.2.4 Cas critique

Une petite remarque pour commencer. $Z_n \to^{p.s.} 0$ mais cette convergence n'est pas dominée par une variable aléatoire L^1 car sinon par convergence dominée, on aurait $\mathbb{E}(Z_n) \to 0$ en décroissance.

Avant de prouver le gros théorème de cette sous-partie, prouvons un lemme qui nous sera utile.

Lemme 7.2.1. *Pour une loi de eproduction μ de moyenne 1 et de variance $\sigma^2 < +\infty$, on a uniformément en s*

$$\frac{1}{n}\left(\frac{1}{1-G_n(s)} - \frac{1}{1-s}\right) \to \frac{\sigma^2}{2}, \text{ où } G_n = G_{Z_n} = G_\mu^{\circ n}$$

Démonstration : $G(1) = 1$, $G'(1) = m = 1$ car on est en Galton-Watson critique, $G''(1) = \mathbb{E}(Z_1^2) - \mathbb{E}(Z_1)$. Par la formule de Taylor avec reste-intégrale

$$G(s) = G(1) + G'(1)(s-1) + (s-1)^2 \int_0^1 G''(1+(s-1)u)(1-u)du$$

$$= G(1)+G'(1)(s-1)+\frac{1}{2}(s-1)^2G''(1)+(s-1)^2 \int_0^1 \{G''(1+(s-1)u)-G''(1)\}(1-u)du$$

Donc,

$$G(s) = 1 + (s-1) + \frac{\sigma^2}{2}(s-1)^2 + (s-1)^2\mathcal{L}(s)$$

avec,

$$\mathcal{L}(s) = \int_0^1 \{G''(1+(s-1)u) - G''(1)\}(1-u)du = \int_0^1 f(s)(1-u)du$$

On a que f est bornée sur $[0,1]$ et $f(s) \to_{s\to 1} 0$.
Par convergence dominée,

$$\mathcal{L}(s) \to_{s\to 1} 0$$

Puis,

$$\frac{1}{1-G(s)} - \frac{1}{1-s} = \frac{G(s)-s}{(1-G(s))(1-s)} = -\frac{\frac{\sigma^2}{2}(s-1)^2 + (s-1)^2\mathcal{L}(s)}{(1-s)(1-s-\frac{\sigma^2}{2}(s-1)^2\mathcal{L}(s))}$$

$$= -\frac{\frac{\sigma^2}{2} + \mathcal{L}(s)}{1-\frac{\sigma^2}{2}(s-1)-(1-s)\mathcal{L}(s)}$$

En résumé

$$\frac{1}{1-G(s)} - \frac{1}{1-s} = \frac{\sigma^2}{2} + \beta(s) \quad (*)$$

où la fonction β est bornée avec $\beta(s) \to_{s\to 1} 0$. Or,

$$\forall k \geq 1,\ G_k(s) \in [0,1[$$

Puis, par $(*)$,

$$\frac{1}{1-G(G_k(s))} - \frac{1}{1-G_k(s)} = \frac{\sigma^2}{2} + \beta(G_k(s))$$

Et, dans le cas critique

$$G_k(s) \to_\infty p_0 = 1$$

Les $(G_k)_k$ converge uniformément vers 1 pour $s \in [0,1[$. Par le théorème de Dini (monotonie + continuité des G_k), $G_n = G \circ G_{n-1}$

$$\frac{1}{n}\left(\frac{1}{1-G(s)} - \frac{1}{1-s}\right) \to_{n\to+\infty} \frac{1}{n}\sum_{k=0}^{n-1}\left(\frac{1}{1-G(G_k(s))} - \frac{1}{1-G_k(s)}\right)$$

Et,

$$\frac{1}{1-G(G_k(s))} - \frac{1}{1-G_k(s)} \to_k \frac{\sigma^2}{2}$$

Si bien que, par le théorème de Cesaro, il vient la convergence uniforme en $s \in [0,1[$ de

$$\frac{1}{n}\left(\frac{1}{1-G(s)} - \frac{1}{1-s}\right) \to \frac{\sigma^2}{2}$$

□

Théorème 7.2.4. $(\sigma^2 < +\infty)$

1. $\lim_{n\to\infty} n\mathbb{P}(Z_n > 0) = \frac{2}{\sigma^2}$;
2. $\lim_{n\to\infty} \frac{1}{n}\mathbb{E}(Z_n | Z_n > 0) = \frac{\sigma^2}{2}$;
3. $\mathcal{L}(n^{-1}Z_n | Z_n > 0) \to^{\mathcal{L}} \mathcal{E}\left(\frac{2}{\sigma^2}\right)$.

On remarque d'ailleurs que

$$\mathbb{E}(n^{-1}Z_n | Z_n > 0) \to \mathbb{E}\left(\mathcal{E}\left(\frac{2}{\sigma^2}\right)\right) = \frac{\sigma^2}{2}$$

Démonstration : Montrons que 1. est vrai. Par le lemme précédent,

$$n\mathbb{P}(Z_n > 0) = n(1 - \mathbb{P}(Z_n = 0)) = n(1 - G_n(0)) = \frac{1}{\frac{1}{n}\left(\frac{1}{1-G_n(0)} - \frac{1}{1-0}\right) + \frac{1}{n}}$$

$$\to \frac{1}{\frac{\sigma^2}{2} + 0} = \frac{2}{\sigma^2}$$

On a ensuite que 1. $\Rightarrow$ 2. car

$$\frac{1}{n}\mathbb{E}(Z_n | Z_n > 0) = \frac{\mathbb{E}(Z_n \mathbb{1}_{Z_n>0})}{n\mathbb{P}(Z_n > 0)} = \mathbb{E}(Z_n)\frac{\sigma^2}{2}$$

Comme on est dans le cas d'un arbre de Galton-Watson critique, $m = 1$ et donc $\mathbb{E}(Z_n) = 1$.

Maintenant, montrons que 3. est vrai. Pour cela, on montre la convergence de la transformée de Laplace pour tout $t > 0$

$$\mathbb{E}(e^{-t\frac{Z_n}{n}}|Z_n > 0) = \frac{\mathbb{E}(e^{-t\frac{Z_n}{n}}\mathbb{1}_{Z_n>0})}{\mathbb{P}(Z_n > 0)} = \frac{\mathbb{E}(e^{-t\frac{Z_n}{n}}) - \mathbb{E}(e^{-t\frac{Z_n}{n}}\mathbb{1}_{Z_n=0})}{1 - G_n(0)}$$

$$= \frac{G_n(e^{-\frac{t}{n}}) - G_n(0)}{1 - G_n(0)} = 1 - \frac{1 - G_n(e^{-\frac{t}{n}})}{1 - G_n(0)}$$

$$= 1 - \frac{1}{n(1 - G_n(0))}\left\{\frac{1}{n}\left(\frac{1}{1 - G_n(e^{-\frac{t}{n}})} - \frac{1}{1 - e^{-\frac{t}{n}}}\right) + \frac{1}{n(1 - e^{-\frac{t}{n}})}\right\}^{-1}$$

Par le lemme précédent

$$\frac{1}{1 - G_n(e^{-\frac{t}{n}})} - \frac{1}{1 - e^{-\frac{t}{n}}} \to \frac{\sigma^2}{2}$$

Or, $n(1 - e^{-\frac{t}{n}}) =_{n\to\infty} n(1 - 1 + \frac{t}{n} + o(\frac{1}{n})) = t + o(1) \to t$.
Donc,

$$\mathbb{E}(e^{-t\frac{Z_n}{n}}|Z_n > 0) \to 1 - \frac{\sigma^2}{2} \times \frac{1}{\frac{\sigma^2}{2} + \frac{1}{t}} = \frac{1}{1 + \frac{t\sigma^2}{2}} = \int_0^\infty e^{-tx}\left(\frac{2}{\sigma^2}e^{-\frac{2}{\sigma^2}x}\right)dx$$

□

7.2.5 Cas sur-critique

Pour un arbre de Galton-Watson sur-critique :

$$m > 1,\ \rho_0 = \mathbb{P}(T < +\infty) < 1,\ Z_n \to^{p.s.} +\infty\mathbb{1}_{\{T=+\infty\}}$$

$$\mathbb{E}(Z_n) = m^n \to +\infty,\ \mathrm{Var}(Z_n) \to +\infty$$

On considère $Y_n = \dfrac{Z_n}{m^n}$. On peut montrer que $(Y_n)_n$ est une martingale.

Théorème 7.2.5. *La martingale $(Y_n)_n$ converge presque sûrement et L^2 vers Y_∞, une variable aléatoire de moyenne 1 et de variance $\dfrac{\sigma^2}{m^2 - m}$.*

Démonstration : On a

$$\mathbb{E}((Y_{n+k} - Y_n)^2) = \mathbb{E}(Y_{n+k}^2 - 2Y_{n+k}Y_n + Y_n^2) = \mathbb{E}(Y_{n+k}^2) - 2\mathbb{E}(\mathbb{E}(Y_{n+k}Y_n|\mathcal{F}_n)) + \mathbb{E}(Y_n^2)$$

Mais,

$$\mathbb{E}(\mathbb{E}(Y_{n+k}Y_n|\mathcal{F}_n)) = \mathbb{E}(Y_n\mathbb{E}(Y_{n+k}|\mathcal{F}_n)) = \mathbb{E}(Y_n^2)$$

Ainsi,

$$\mathbb{E}((Y_{n+k} - Y_n)^2) = \mathbb{E}(Y_{n+k}^2) - \mathbb{E}(Y_n^2) = \frac{\sigma^2 m^{n+k}(m^{n+k} - 1)}{m^{2n+2k}(m^2 - m)} - \frac{\sigma^2 m^n(m^n - 1)}{m^{2n}(m^2 - m)}$$

$$= \frac{\sigma^2}{m^n} \frac{1 - m^{-k}}{m^2 - m}, \ (*)$$

Donc

$$\sup_{k \in \mathbb{N}} \mathbb{E}((Y_{n+k} - Y_n)^2) \to_{n \to \infty} 0$$

Par le critère de Cauchy, on a la convergence dans L^2 vers une variable L^2 et on en déduit l'espérance et la variance en passant à la limite dans $(*)$.

$$\mathbb{E}((Y_n - Y_\infty)^2) = \frac{\sigma^2}{m^n(m^2 - m)} = \mathcal{O}_{n \to \infty}\left(\frac{1}{m^n}\right)$$

Donc, par l'inégalité de Markov

$$\mathbb{P}(|Y_n - Y_\infty| \geq \epsilon) \leq \frac{\mathbb{E}(|Y_n - Y_\infty|^2)}{\epsilon^2} = \mathcal{O}_{n \to \infty}\left(\frac{1}{m^n}\right)$$

Ainsi, par le lemme de Borel-Cantelli

$$\mathbb{P}(\limsup_n |Y_n - Y_\infty| > \epsilon) = 0$$

C'est-à-dire que presque sûrement, on a

$$\liminf |Y_n - Y_\infty| < \epsilon = \bigcup_k \bigcap_{n > k} \{|Y_n - Y_\infty < \epsilon\}$$

Pour tout $\epsilon > 0$, il existe $A = A(\epsilon) \in \mathcal{F}$ de mesure 1 tel que sur A

$$\exists k, \ \forall n \geq k, \ |Y_n - Y_\infty| < \epsilon$$

Donc,

$$\limsup |Y_n - Y_\infty| < \epsilon$$

Prenons, $\epsilon = \frac{1}{p}$. Les évènements $A(\frac{1}{p})$ sont de probabilité 1.

On pose $\bigcap_p A\left(\frac{1}{p}\right)$ de probabilité 1.

$$\forall p \geq 1, \limsup_n |Y_n - Y_\infty| < \frac{1}{p}$$

Donc,

$$\limsup_n |Y_n - Y_\infty| = 0$$

Si bien que

$$Y_n \to^{p.s.} Y_\infty$$

□

Théorème 7.2.6. *La transformée de Laplace $L_\infty(s) = \mathbb{E}(e^{-sY_\infty})$ vérifie l'équation*

$$L_\infty^{'}(0) = 1$$

$$L_\infty(ms) = G(L_\infty(s))$$

Démonstration : On a

$$L_\infty(s) = \mathbb{E}(e^{-sY_\infty}) \text{ et } L_\infty^{'}(0) = \mathbb{E}(Y_\infty) = 1$$

$$L_n(s) = \mathbb{E}(e^{-sY_n}) = \mathbb{E}(e^{-s\frac{Z_n}{m^n}}) = G_n(e^{-\frac{s}{m^n}})$$

$$L_{n+1}(s) = G(G_n(e^{-\frac{sm}{m^{n+1}}})) = G(L_n(s)) \to G(L_\infty(s))$$

□

7.2.6 Résumé sur les différents cas

Arbre sous-critique : $m < 1$

- $\mathbb{E}(Z_n) = m^n \to 0$ en décroissance, i.e. la population s'éteint avec probabilité 1 ;
- L'arbre est presque sûrement fini ;
-

Arbre critique : $m = 1$

- $\mathbb{E}(Z_n) = 1$ pour tout n, la population s'éteint aussi presque sûrement ;
-

Arbre sur-critique : $m > 1$

- $\mathbb{E}(Z_n) \to +\infty$. Il y a survie de l'arbre avec probabilité non-nulle : La probabilité d'extinction $\rho_0 < 1$, solution de $\rho_0 = G(\rho_0)$;
-

7.2.7 Immigration

On suppose qu'à chaque génération arrivent des individus extérieurs pour contribuer à former la nouvelle génération.
En notant Z_n^I la taille de cette population à la date n, on a

$$Z_{n+1}^I = \sum_{k=1}^{Z_n^I} X_{n+1,k} + I_{n+1}$$

où, les $X_{n+1,k}$ sont indépendantes et identiquement distribuées de loi de reproduction μ, et I_{n+1} est le nombre de nouveayx individus de loi d'immigration μ_+.
On suppose que $X_{n,k}$, $n \geq 1$, $k \geq 1$, I_n $\forall n$, Z_0 forment une famille de variables aléatoires indépendantes. On note de plus m_+ la moyenne de μ_+ ; σ_+^2 la variance de μ_+ ; H la fonction génératrice.
Dans la suite

$$\mathcal{F}_n = \sigma(Z_0 = 1, (X_{m,k})_{m \leq n, k \geq n}, (I_j)_{j \leq m})$$

$$\mathcal{G} = \sigma(I_n | n \geq 1) \text{ tribu de l'immigration}$$

On suppose $m_+ < \infty$, on déduit la taille moyenne de la population :

$$\mathbb{E}(Z^I_{n+1}) = \mathbb{E}\Big(\mathbb{E}\Big(\sum_{k=1}^{Z^I_n} X_{n+1,k} + I_{n+1}\Big|\mathcal{F}_n\Big)\Big) = \mathbb{E}\Big(\sum_{k=1}^{Z^I_n} \mathbb{E}(X_{n+1,k}) + \mathbb{E}(I_{n+1})\Big) = \mathbb{E}\Big(\sum_{k=1}^{Z^I_n} m + m_+\Big)$$

car, $X_{n+1,k}, I_{n+1} \perp\!\!\!\perp \mathcal{F}_n$. On en déduit donc

$$E(Z^I_{n+1}) = m\mathbb{E}(Z^I_n) + m_+$$

D'où,

$$\mathbb{E}(Z^I_n) = \Big(m^n + \frac{m^n - 1}{m-1} m_+\Big)\mathbb{1}_{m\neq 1} + (1 + nm_+)\mathbb{1}_{m=1}$$

On peut aussi calculer $\mathrm{Var}(Z^I_n)$ en supposant $\sigma^2 < \infty$, i.e. lorsque μ et μ_+ admettent des moments d'ordres 2 :

$$\mathbb{E}((Z^I_{n+1})^2|\mathcal{F}_n) = \sigma^2 Z_n + m^2 Z_n + 2mm_+ Z_n + (\sigma^2_+ + m^2_+)$$

Par récurrence

$$\mathrm{Var}(Z^I_{n+1}) = \mathbb{E}((Z^I_{n+1})^2) - \mathbb{E}(Z^I_{n+1})^2 = m^2 \mathrm{Var}(Z^I_n) + \sigma^2 \mathbb{E}(Z^I_n) + \sigma^2_+$$

On suppose $Z_0 = 1$ et on note G^I_n la fonction génératrice de Z^I_n.

$$G^I_{n+1}(s) = \mathbb{E}\big(s^{Z^I_{n+1}}\big) = \mathbb{E}\Big(\prod_{k=1}^{Z^I_n} s^{X_{n+1,k}} s^{In+1}\Big) =^{\perp\!\!\!\perp} \mathbb{E}\Big(\prod_{k=1}^{Z^I_n} s^{X_{n+1,k}}\Big)\mathbb{E}\big(s^{I_{n+1}}\big) = \mathbb{E}\Big(\prod_{k=1}^{Z^I_n} s^{X_{n+1,k}}\Big) H(s)$$

Puis,

$$\mathbb{E}\Big(\prod_{k=1}^{Z^I_n} s^{X_{n+1,k}}\Big) = \mathbb{E}\Big(\mathbb{E}\Big(\prod_{k=1}^{Z^I_n} s^{X_{n+1,k}}\Big|\mathcal{F}_n\Big)\Big) = \mathbb{E}\Big(\prod_{k=1}^{Z^I_n} \mathbb{E}(s^{X_{n+1,k}})\Big) = \mathbb{E}\Big(\prod_{k=1}^{Z^I_n} G(s)\Big) = \mathbb{E}\big(G(s)^{Z^I_n}\big)$$

C'est-à-dire,

$$\mathbb{E}\Big(\prod_{k=1}^{Z^I_n} s^{X_{n+1,k}}\Big) = G^I_n(G(s))$$

On a donc

$$G^I_{n+1}(s) = H(s) G^I_n(s)$$

On calcule pour certain n :

$$G^I_0(s) = s, \text{ car } Z^I_0 = 1$$
$$G^I_1(s) = G(s)H(s), \text{ car } Z^I_1 = X_{1,1} +^{\perp\!\!\!\perp} I_1$$
$$G^I_2(s) = H(s)G^I_1(G(s)) = H(s)H(G(s))G_2(s)$$
$$G^I_3(s) = H(s)G^I_2(G(s)) = H(s)H(G(s))H(G_2(s))G_3(s)$$

D'où on déduit par récurrence,

$$G^I_n(s) = H(s)H(G(s))...H(G_{n-1}(s))G_n(s)$$

Exemple 7.5. On pose

$$\mu = \mathcal{B}(p) = (1-p)\delta_0 + p\delta_1$$
$$\mu_+ = \mathcal{P}(\lambda)$$

Alors, on a

$$G(s) = (1-p) + ps \text{ et } H(s) = e^{\lambda(s-1)}$$

Puis,

$$G_2(s) = G(G(s)) = 1 - p + pG(s) = 1 - p + p(1 - p + ps) = 1 - p^2 + p^2 s$$

Par récurrence,

$$G_n(s) = 1 - p^n + p^n s$$

Maintenant,

$$G_n^I(s) = G_n(s) \prod_{k=0}^{n-1} H(G_k(s)) = (1 - p^n + p^n s) \prod_{k=0}^{n-1} \exp\left(\lambda(1 - p^k + p^k s - 1)\right)$$

$$= (1 - p^n + p^n s) \exp\left(\sum_{k=0}^{n-1} \lambda p^k (s-1)\right) = (1 - p^n + p^n s) \exp\left(\lambda \frac{1-p^n}{1-p}(s-1)\right)$$

$$\to \exp\left(\frac{\lambda}{1-p}(1-s)\right) \sim \mathcal{P}\left(\frac{\lambda}{1-p}\right)$$

7.2.8 Arbre de Galton-Watson multiple

7.3 Processus de Poisson

7.3.1 Rappels probabilistes

Loi de Poisson :

Soit $X \sim \mathcal{P}(\lambda)$ à support dans $\mathbb{N}$, alors

$$\mathbb{P}(X = n) = \frac{\lambda^n}{n!} e^{-\lambda},\ \mathbb{E}(X) = \lambda$$

$$\varphi_X(t) = \mathbb{E}(e^{itX}) = e^{\lambda(e^{it}-1)},\ G_X(t) = \mathbb{E}(t^X) = e^{\lambda(t-1)}$$

Proposition 7.3.1 (Superposition)**.** *Soit $X \sim \mathcal{P}(\lambda)$ et $Y \sim \mathcal{P}(\mu)$, avec $X \perp\!\!\!\perp Y$. Alors $X + Y \sim \mathcal{P}(\lambda + \mu)$.*

Démonstration : Notons $Z = X + Y$. De plus, notons $\sum a_n t^n$, $\sum b_n t^n$, $\sum c_n t^n$, les génératrices respectives de X, Y et Z. Alors par hypothèses,

$$a_n = \frac{\lambda^n}{n!} e^{-\lambda},\ b_n = \frac{\mu^n}{n!} e^{-\mu}$$

Il suffit donc de calculer c_n via Cauchy,

$$c_n = \sum_{k=0}^{n} a_k b_{n-k} = e^{-(\lambda+\mu)} \sum_{k=0}^{n} \frac{\lambda^k \mu^{n-k}}{k!(n-k)!} = \frac{e^{-(\lambda+\mu)}}{n!} \sum_{k=0}^{n} \binom{n}{k} \lambda^k \mu^{n-k} = \frac{(\lambda+\mu)^n}{n!} e^{-(\lambda+\mu)}$$

□

Proposition 7.3.2 (Amincissement). *Soit $X \sim \mathcal{P}(\lambda)$, $(\varepsilon_i)_{i\geq 1}$ i.i.d. de loi $\mathcal{B}(p)$ où $X \perp\!\!\!\perp \varepsilon_i$. Alors*

$$Y = \sum_{i=1}^{X} \varepsilon_i \sim \mathcal{P}(\lambda p)$$

Démonstration : On calcule usuellement la génératrice de Y qui caractérise la loi

$$G_Y(t) = \mathbb{E}(t^Y) = \mathbb{E}(t^{\sum_{i=1}^{X}\varepsilon_i}) = \mathbb{E}(\mathbb{E}(t^{\sum_{i=1}^{X}\varepsilon_i}|X)) = \mathbb{E}\bigg(\mathbb{E}\bigg(\prod_{i=1}^{X} t^{\varepsilon_i}\bigg|X\bigg)\bigg)$$

Par indépendance,

$$= \mathbb{E}\bigg(\prod_{i=1}^{X}\mathbb{E}(t^{\varepsilon_i})\bigg) = \mathbb{E}\bigg(\prod_{i=1}^{X}\{(1-p)+pt\}\bigg) = \mathbb{E}(\{(1-p)+pt\}^X) = G_X(1-p+pt) = e^{\lambda p(t-1)}$$

□

Le fait qu'en sommant des lois de Bernoulli on obtienne une loi de Poisson vient du fait que dans la somme, le nombre d'évènements ε_i est aussi aléatoire. Si on avait un nombre d'évènement fixe, on obtiendrait une loi binomiale comme d'habitude.

Définition 7.3.1 (Loi de Poisson composée). On appelle loi de Poisson composée la loi de

$$Z = \sum_{i=1}^{N} X_i$$

où $N \sim \mathcal{P}(\lambda)$ et $N \perp\!\!\!\perp X_i$, (X_i) indépendants et de même loi.

Proposition 7.3.3. *Pour Z une variable aléatoire de Poisson composée*

1. $\varphi_Z(t) = e^{\lambda(\varphi_X(t)-1)}$;
2. *Si $X_i \in L^1$, alors $Z \in L^1$ et $\mathbb{E}(Z) = \lambda\mathbb{E}(X_1)$;*
3. *Si $X \in L^2$ alors $Z \in L^2$ et $Var(Z) = \lambda\mathbb{E}(X_1^2)$.*

Démonstration : Pour la première affirmation, par indépendance on a

$$\varphi_Z(t) = \mathbb{E}\bigg(\mathbb{E}\bigg(e^{it\sum_{k=1}^{N}X_k}\bigg|N\bigg)\bigg) = \mathbb{E}\bigg(\mathbb{E}\bigg(\prod_{k=1}^{N}e^{itX_k}\bigg|N\bigg)\bigg) = \mathbb{E}\big((\varphi_X(t))^N\big)$$

$$= e^{\lambda(\varphi_X(t)-1)}$$

Puis, pour la deuxième affirmation, on a de même

$$\mathbb{E}(Z) = \mathbb{E}\bigg(\mathbb{E}\bigg(\sum_{k=1}^{N}X_k\bigg|N\bigg)\bigg) = \mathbb{E}\big(N\mathbb{E}(X_1)\big) = \lambda\mathbb{E}(X_1)$$

□

Théorème 7.3.1 (de Raikov). *Soit $X = Y + Z$ avec $Y \perp\!\!\!\perp Z$ positives. Alors X est une loi de Poisson si et seulement si Y et Z le sont.*

Démonstration : Le sens réciproque est immédiat par ce qui précéde.
Voyons le sens direct. Supposons que $X = Y + Z$ est de Poisson, avec $Y \perp\!\!\!\perp Z$ positives et montrons que Y et Z sont de loi de Poisson.
□

Loi exponentielle :

Soit $X \sim \mathcal{E}(\lambda)$. X est de support dans $\mathbb{R}_+$ et de densité

$$f(x) = \lambda e^{-\lambda x} \mathbb{1}_{x \geq 0}$$

De plus, on a par intégration

$$\mathbb{E}(X) = \frac{1}{\lambda}, \ \mathrm{Var}(X) = \frac{1}{\lambda^2}$$

$$\varphi_X(t) = \frac{\lambda}{\lambda - it}$$

$$\mathbb{P}(X > t) = e^{-\lambda t}, \ t > 0$$

Proposition 7.3.4 (Absence de mémoire). *Soit X une variable aléatoire positive. Alors, X est de loi exponentielle si et seulement si*

$$\mathbb{P}(X > t + s | X > s) = \mathbb{P}(X > t), \ \forall t, s > 0$$

Démonstration : Le sens direct est assez immédiat

$$\mathbb{P}(X > t+s | X > s) = \frac{\mathbb{P}(X > t + s, X > s)}{\mathbb{P}(X > s)} = \frac{\mathbb{P}(X > t + s)}{\mathbb{P}(X > s)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t} = \mathbb{P}(X > t)$$

Quand au sens réciproque, on utilise notre fonction de survie

$$\overline{F}(t) = 1 - F(t) = \mathbb{P}(X > t)$$

On a

$$\overline{F}(t + s) = \overline{F}(t)\overline{F}(s)$$

Or, $\overline{F}(0) = 1$ car X est positive. Après calcul,

$$\overline{F}(t) = e^{-\lambda t}$$

□

Proposition 7.3.5. *Soient $X_i \sim \mathcal{E}(\alpha_i)$ indépendantes. On pose*

$$Z = \min_i X_i, \quad K = Argmin_i X_i$$

Alors,

$$Z \sim \mathcal{E}\left(\sum_{i=1}^{n} \alpha_i \right) \perp\!\!\!\perp K$$

On remarque aussi que les lois exponentielles sont des cas particuliers de la loi Γ.

Définition 7.3.2 (Loi Gamma)**.** On note $\Gamma(p, \lambda)$ avec $p, \lambda > 0$ la loi Gamma de densité

$$f(x) = \frac{\lambda^p}{\Gamma(p)} x^{p-1} e^{\lambda x} \mathbb{1}_{\mathbb{R}_+}(x)$$

On rappelle que la fonction Γ est définie pour tout $z \in \mathbb{C}$ tel que $\mathrm{Re}(z) > 0$ et s'exprime comme

$$\Gamma(z) = \int_0^{+\infty} t^{z-1} e^{-t} dt$$

Concernant la fonction Γ, on montre assez aisément que $\Gamma(1) = 1$ et que $\Gamma(z+1) = z\Gamma(z)$.

On a donc les relations suivantes sur la loi,

$$\mathcal{E}(\alpha) = \Gamma(1, \alpha)$$

$$\Gamma(p, \alpha) * \Gamma(q, \alpha) = \Gamma(p+q, \alpha)$$

$$\mathcal{E}(\alpha)^{*n} = \Gamma(n, \alpha)$$

$$\chi^2(p) = \Gamma\left(\frac{p}{2}, \frac{1}{2}\right)$$

7.3.2 Notions de processus en temps continu

Définition 7.3.3 (Processus stochastique)**.** Un processus stochastique $(X_t)_{t \in T}$ est une famille de variable aléatoire indexée par un ensemble T.

Passons en revu plusieurs choix de T :

- $T = \{t_0\}$: Variable aléatoire ;
- $T = \{t_0, ..., t_n\}$ fini : Vecteur aléatoire ;
- $T = \mathbb{N}$: Suite de variable aléatoire ;
- $T = [0, +\infty[$: Cadre usuel pour lequel $t \in T$ s'interprète comme le temps ;
- $T \subseteq \mathbb{R}^d$: Champ aléatoire (Drap pour $d = 2$).

Pour la suite, on prend $T = \mathbb{R}_+$. Dans ce cadre, un processus stochastique $(X_t)_{t \geq 0}$ est difficile car $\mathbb{R}_+$ n'est pas dénombrable.
Pour lever cet obstacle, on va supposer une hypothèse de régularité de la trajectoire $t \mapsto X_t(\omega)$, $\forall \omega \in \Omega$, du type

- Continuité ;
- Continue à droite limite à gauche (càdlàg)

$$\lim_{s \to t^+} X_s(\omega) = X_t(\omega), \ \lim_{s \to t^-} X_s(\omega) \text{ existe}$$

Dans ce cadre, on donne un sens à des évènements du type

$$\{X_s = x | \forall s \leq t\} = \bigcap_{s \leq t,\ s \in \mathbb{Q}} \{X_s = x\}$$

Pour des processus stochastiques $(X_t)_{t \geq 0}$ on associe encore la filtration canonique

$$\mathcal{F}_t = \sigma(X_s | s \leq t), \text{ contient l'information avant } t$$

$$= \sigma\left(\bigcup_{s \leq t} \sigma(X_s) \right)$$

On rappelle que

$$s \leq t \Rightarrow \mathcal{F}_s \subseteq \mathcal{F}_t$$

Définition 7.3.4 (Temps d'arrêt)**.** On définit un temps d'arrêt T à valeurs dans $[0, \infty[\cup \{+\infty\}$ par la propriété

$$\forall t \in \mathbb{R}_+,\ \{t \leq t\} \in \mathcal{F}_t$$

Remarque 7.2.

$$\{T \leq t\} \in \mathcal{F}_t \ : \ \bigcup_{n \in \mathbb{N}^*} \left\{ T \leq t - \frac{1}{n} \right\} \subseteq \{T < t\}$$

$$\left\{ T \leq t - \frac{1}{n} \right\} \in \mathcal{F}_{t - \frac{1}{n}} \subseteq \mathcal{F}_t$$

$$T < t \Rightarrow \exists N \geq 1 \mid T \leq t - \frac{1}{n}$$

Définition 7.3.5 (Tribu du temps d'arrêt)**.** A un temps d'arrêt T, on associe la tribu $\mathcal{F}_T$

$$\mathcal{F}_T = \left\{ A \in \mathcal{F} \middle| A \cap \{T \leq t\} \in \mathcal{F}_t, \forall t \geq 0 \right\}$$

7.3.3 Processus de comptage

Le processus de Poisson est un processus de comptage d'évènements qui se réalisent successivement au cours du temps. Par exemple, l'arrivée des appels à un serveur téléphonique, les files d'attentes, dates de sinistres qui frappent un assuré...

Dans la suite, on s'intéresse à des évènements de référence et on note les dates d'occurence de ces évènements comme suit :

$$T_0 = 0 < T_1 < T_2 < ... < T_n < ...$$

On suppose que presque sûrement

$$\lim_{n \to \infty} T_n = +\infty$$

On pose alors

$$N_t = \sup\{n | T_n \leq t\}$$

(Nombre d'évènements de référence qui ont lieu jusqu'à t)

On remarque que N_t est càdlàg.

$$(N_t)_{t\geq 0} \longleftrightarrow \tau = (T_n)$$

Processus de comptage $\longleftrightarrow$ Processus ponctuel

On remarque,

$$T_n = \inf\{t | N_t = n\}$$

- $N_t = n \Longleftrightarrow t \geq T_n \wedge t < T_{n+1}$;
- $N_t \geq n \Longleftrightarrow t \geq T_n$;
- $N_s < n \leq N_t \Longleftrightarrow T_n \in]s, t[$.

$$N_t = \sum_{k\geq 1} \mathbb{1}_{T_k \leq t} \text{ (comptage)}$$

On définit maintenant formellement l'élément central de cette section,

Définition 7.3.6 (Processus de Poisson)**.** Un processus de comptage $(N_t)_{t\geq 0}$ est un processus de Poisson d'intensité $\lambda > 0$ si :

1. $N_0 = 0$, $t \mapsto N_t$ càdlàg ;
2. Accroissements indépendants :
   Pour tout $0 \leq t_1 < t_2 < ... < t_k$, les variables aléatoires N_{t_1}, $N_{t_2} - N_{t_1}$,...,$N_{t_k} - N_{t_{k-1}}$ sont indépendants ;
3. Accroissements stationnaires :
   Pour tout $0 \leq s \leq t$, $\mathcal{L}(N_t - N_s)$ ne dépend que de $\Delta = t - s$
4. Bernoulli infinitésimal :

$$\mathbb{P}(N_t = k) = \begin{cases} 1 - \lambda t + o(t), \ k = 0 \\ \lambda t + o(t), \ k = 1 \\ o(t), \ k = 2 \end{cases}$$

On remarque que pour $t << 1$, $N_t \approx \mathcal{B}(\lambda t)$.

- Pour 3., $\mathcal{L}(N_t - N_s) \sim \mathcal{L}(N_{t-s} - N_0(= 0)) = \mathcal{L}(N_{t-s})$;
- 4. peut être montré via 2. et 3.

Proposition 7.3.6. *Pour un processus de Poisson d'intensité $\lambda > 0$, on a*

$$N_t \sim \mathcal{P}(\lambda t)$$

Démonstration : Montrons que pour tout t, $N_t \sim \mathcal{P}(\lambda t)$. Soit

$$\mathbb{P}_k(t) = \mathbb{P}(N_t = k)$$

Alors,

$$N_t \sim \mathcal{P}(\lambda t) \Leftrightarrow \mathbb{P}_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \Leftrightarrow G_t(x) = \mathbb{E}(x^{N_t}) = e^{\lambda t(x-1)}$$

On a,

$$\mathbb{P}_0(t+h) = \mathbb{P}(N_{t+h} = 0) = \mathbb{P}(N_{t+h} - N_t = 0, N_t = 0) =^{\perp\!\!\!\perp} \mathbb{P}(N_h = 0)\mathbb{P}(N_t = 0)$$

$$= (1 - \lambda h + o(h))\mathbb{P}_0(t)$$

Donc,

$$\frac{\mathbb{P}_0(t+h) - \mathbb{P}_0(t)}{h} = -\lambda \mathbb{P}_0(t) + o(1)$$

On prend l'asymptotique $[h \to 0]$ et il vient que

$$\frac{d\mathbb{P}_0}{dt}(t) = \mathbb{P}_0^{'}(t) = -\lambda \mathbb{P}_0(t)$$

Or,

$$\mathbb{P}_0(0) = \mathbb{P}(N_0 = 0) =^{\Delta} 1$$

D'où,

$$\mathbb{P}_0(t) = e^{-\lambda t}$$

Maintenant, soit $k \geq 1$,

$$\mathbb{P}_k(t+h) = \mathbb{P}(N_{t+h} = k) = \sum_{j=0}^{k} \mathbb{P}(N_{t+h} = k | N_t = j)\mathbb{P}(N_t = j)$$

$$= \sum_{j=0}^{k} \mathbb{P}(N_{t+h} - N_t = k - j | N_t - N_0 = j)\mathbb{P}(N_t = j)$$

$$= \sum_{j=0}^{k} \mathbb{P}(N_{t+h} - N_t = k - j)\mathbb{P}(N_t = j) = \sum_{j=0}^{k} \mathbb{P}(N_h = k - j)\mathbb{P}(N_t = j)$$

$$= \mathbb{P}_k(t)\mathbb{P}(N_h = 0) + \mathbb{P}_{k-1}(t)\mathbb{P}(N_h = 1) + \sum_{j=0}^{k-2} \mathbb{P}(N_h = k - j)\mathbb{P}(N_t = j)$$

Or, $\mathbb{P}(N_h = 0) = 1 - \lambda h + o(h)$ et $\mathbb{P}(N_h = 1) = \lambda h + o(h)$ et $\mathbb{P}(N_h = k - j) = o(t)$ pour $j = 0, ..., k-2$. Donc,

$$\Longleftrightarrow \frac{\mathbb{P}_k(t+h) - \mathbb{P}_k(t)}{h} = -\lambda \mathbb{P}_k(t) + \lambda \mathbb{P}_{k-1}(t) + o(1)$$

D'où en prenant l'asymptotique $[h \to 0]$

$$\mathbb{P}_k^{'}(t) = -\lambda(\mathbb{P}_k(t) - \mathbb{P}_{k-1}(t)),\ k \geq 1 \quad (*)$$

On suppose vrai que : $\mathbb{P}_{-1}(t) = 0$. Pat $(*)$ il vient que,

$$\sum_{k\geq 0} \mathbb{P}_k^{'}(t)x^k = -\lambda \sum_{k\geq 0} \mathbb{P}_k(t)x^k - \mathbb{P}_{k-1}(t)x^k$$

$$\Leftrightarrow \sum_{k\geq 0} \mathbb{P}_k^{'}(t)x^k = -\lambda \sum_{k\geq 0}(t)x^k + \lambda x \sum_{k\geq 1} \mathbb{P}_{k-1}(t)x^{k-1} = -\lambda G_t(x) + \lambda G_t(x)$$

$$\Leftrightarrow \sum_{k\geq 0} \mathbb{P}_k^{'}(t)x^k = \lambda G_t(x)(x-1)$$

Ainsi, $\sum_{k\geq 0} \mathbb{P}_k^{'}(t)x^k$ converge normalement en t. Si bien que,

$$\sum_{k=0}^{\infty} \mathbb{P}_k^{'}(t)x^k = \sum_{k=0}^{\infty} \frac{d\mathbb{P}_k(t)}{dt}x^k = \frac{d}{dt}G_t(x)$$

D'où,

$$\frac{\partial}{\partial t}G_t(x) = \lambda G_t(x)(x-1)$$

Puis,

$$G_t(x) = \gamma e^{\lambda(x-1)t}$$

Or, $N_0 = 0$ et donc $G_0(x) = \mathbb{E}(x^{N_0}) = 1$. Finalement,

$$G_t(x) = e^{\lambda(x-1)t}$$

□

Conséquences :

1. $\mathbb{E}(N_t) = \lambda t \Rightarrow \lambda = \frac{\mathbb{E}(N_t)}{t}$. Nombre moyen de référence par unité de temps.
2. Le processus de Poisson n'a pas de saut fixe, i.e.

   $$\forall t > 0 \text{ fixé },\ \Delta_t = N_t - N_{t^-} = 0 \text{ p.s.}$$

   En effet,

   $$\Delta_t = \lim_{h\to 0}(N_t - N_{t-h}) \sim N_h \sim \mathcal{P}(\lambda h) \to^{\mathcal{L}} \delta_0$$

   Donc, $\Delta_t \sim^{\mathcal{L}} \delta_0$, i.e. $\Delta_t = 0$ p.s.
3. Soit $0 \leq t_0 \leq t_1 \leq ... \leq t_n$ alors les accroissements $N_{t_1} - N_{t_0}$,...,$N_{t_n} - N_{t_{n-1}}$ sont indépendants et de loi

   $$\mathcal{P}(\lambda(t_1 - t_0)), ..., \mathcal{P}(\lambda(t_n - t_{n-1}))$$

   D'où la loi de $(N_{t_0}, ..., N_{t_n})$.

On va maintenant voir 2 théorèmes importants de cette théorie.

Théorème 7.3.2 (de Markov-faible)**.** *Soit $(N_t)_{t>0}$ un processus de Poisson d'intensité $\lambda > 0$. Alors,*

$$\forall s > 0 \text{ fixé }, \ N^{(s)} = N_{t+s} - N_s, \ t > 0$$

est un processus de Poisson d'intensité λ et $N^{(s)}$ est indépendant de $\mathcal{F}^s = \sigma(N_t | t \leq s)$.

Démonstration : Plus tard. □

Théorème 7.3.3 (de Markov-fort)**.** *Le même énoncé est vrai en remplaçant le temps déterministe s par une variable aléatoire S qui est un $(\mathcal{F}_t)$-temps d'arrêt.*
En particulier, $N^{(S)} \perp\!\!\!\perp \mathcal{F}_S$ (tribu associée aux temps d'arrêt).

7.3.4 Structures de sauts

On rappelle que T_n est la date du saut n ($T_0 = 0$) et que $\Delta_n = T_n - T_{n-1}$ est la durée du n-ième inter-saut.

Proposition 7.3.7.

7.3.5 Caractérisation d'un processus de Poisson

Théorème 7.3.4 (de caractérisation)**.** *Soit $(N_t)_{t\geq 0}$ une famille de variables aléatoires dans $\mathbb{N}$. On suppose que $N_0 = 0$ et que $t \mapsto N_t$ est càdlàg. Alors les propriétés suivantes sont équivalentes et définissent un processus de Poisson d'intensité $\lambda > 0$:*

1. ***Comptage :*** *$(N_t)_{t\geq 0}$ est un processus de comptage d'évènements de référence espacés d'une durée indépendantes et de même loi $\mathcal{E}(\lambda)$.*
2. ***Sauts :*** *Les durées entre les sauts de $(N_t)_{t\geq 0}$ sont indépendantes et de même loi $\mathcal{E}(\lambda)$ et les sauts valent ± 1.*
3. ***Structure des accroissements :*** *Pour tout $t_0 = 0 < t_1 < ... < t_n$, les accroissements $N_{t_1} - N_{t_0}, ..., N_{t_n} - N_{t_{n-1}}$ sont indépendants et de loi $\mathcal{P}(\lambda(t_{i+1} - t_i))$.*
4. ***Propriété infinitésimale :*** *Les accroissements de $(N_t)_{t\geq 0}$ sont indépendants et*

$$\sup_{t\geq 0} \mathbb{P}(N_{n+h} - N_t = k) = \begin{cases} 1 - \lambda h + o(h), \ k = 0 \\ \lambda h + o(h), \ k = 1 \\ o(h), \ k = 2 \end{cases}$$

Démonstration : La première équivalence est immédiate. On verra le reste plus tard. □

Théorème 7.3.5. *Le processus de Poisson est le seul processus de comptage simple (évènements non-simultanés) à accroissements indépendants et stationnaires.*

Démonstration : Long. □

7.3.6 Opérations sur les processus de Poisson

Théorème 7.3.6 (Superposition). *Si $(N_t^{(\lambda_1)})_{t>0}$ et $(N_t^{(\lambda_2)})_{t>0}$ sont des processus de Poisson d'intensité λ_1 et λ_2, et indépendants alors*

$$N_t = N_t^{(\lambda_1)} + N_t^{(\lambda_2)}$$

est un processus de Poisson d'intensité $\lambda_1 + \lambda_2$.

Démonstration : On pose alors

$$N_t = N_t^{(\lambda_1)} + N_t^{(\lambda_2)}$$

Comme les deux éléments de la décomposition sont des processus de Poisson, alors d'une part

$$N_0 = N_0^{(\lambda_1)} + N_0^{(\lambda_2)} = 0 + 0 = 0$$

D'autre part, les $t \mapsto N_t^{(\lambda_i)}$ sont càdlàg donc $t \mapsto N_t$ l'est aussi.
Maintenant, soit $n \geq 1$, $0 < t_0 < t_1 < ... < t_n$, on a

$$N_{t_i} - N_{t_{i-1}} = (N_{t_i}^{(\lambda_1)} - N_{t_{i-1}}^{(\lambda_1)}) + (N_{t_i}^{(\lambda_2)} - N_{t_{i-1}}^{(\lambda_2)})$$

et,

$$(N_{t_i}^{(\lambda_1)} - N_{t_{i-1}}^{(\lambda_1)}) \perp\!\!\!\perp (N_{t_i}^{(\lambda_2)} - N_{t_{i-1}}^{(\lambda_2)})$$

Si bien que,

$$(N_{t_i}^{(\lambda_1)} - N_{t_{i-1}}^{(\lambda_1)}) \perp\!\!\!\perp (N_{t_j}^{(\lambda_1)} - N_{t_{j-1}}^{(\lambda_1)}) \perp\!\!\!\perp (N_{t_j}^{(\lambda_2)} - N_{t_{j-1}}^{(\lambda_2)})$$

Donc,

$$N_{t_i} - N_{t_{i-1}} \perp\!\!\!\perp N_{t_j} - N_{t_{j-1}}$$

Puis,

$$N_{t+s} - N_t = \Delta N^{(\lambda_1)} + \Delta N^{(\lambda_2)} \sim N_t^{(\lambda_1)} + N_t^{(\lambda_2)} \sim N_t$$

Finalement,

$$\mathbb{E}(N_t) = \lambda_1 t + \lambda_2 t = (\lambda_1 + \lambda_2)t$$

□

Passons maintenant à une autre opération. L'amincissement d'un processus de Poisson (ponctuel) est l'opération qui consiste à retirer des sauts indépendamment les uns des autres avec probabilité $1 - p$.

Soit $(\varepsilon_k)_{k\geq 0} \sim \mathcal{B}(p)$ indépendants et de même loi, $\perp\!\!\!\perp (N_t)_{t\geq 0}$. Le processus p-aminci est

$$N_t^{(p)} = \sum_{k\geq 1} \mathbb{1}_{\{T_k \leq t\} \cap \{\varepsilon_k = 1\}}$$

En gros, la suite (ε_k) nous sélectionne les sauts. On saute que $\varepsilon_k = 1$ et on ne saute pas quand $\varepsilon_k = 0$.

Théorème 7.3.7 (Amincissement). *$N^{(p)}$ est un processus d'intensité λp. De plus,*

$$N^{(q)} = N - N^{(p)}$$

est un processus $(1-p)$-aminci et

$$N^{(q)} = N - N^{(p)} \perp\!\!\!\perp N^{(p)}$$

Démonstration : $N^{(p)}$ est un processus de comptage. On détermine les durées inter-sauts de $N^{(p)}$ en notant T_k' la date du k-ième saut de $N^{(p)}$

7.4 Processus de Markov de saut

7.4.1 Chaîne de Markov en temps continu

7.4.2 Taux de transition

7.4.3 Durée de séjour

7.4.4 Equations de Kolmogorov

7.4.5 Propriété des chaînes de Markov de saut pur

7.5 Processus de naissance et de mort

7.5.1 Généralités

7.5.2 Files d'attentes

Le phénomène de file d'attente apparait dès qu'un service est proposé à des utilisateurs qui doivent possiblement attendre avant d'y accéder. Leur modélisation remonte à Erlang vers 1900 à Copenhague avec une approche markovienne. Les files d'attentes appartiennent plus largement à la branche *Operating System* des Mathématiques. Il s'agit de cas particuliers de processus de naissance et de mort.
Dans la suite, on note :

- X_t le nombre d'utilisateurs dans le système à la date $t \in \mathbb{R}_+^*$;
- Arrivées : $T_1, T_2, ..., T_n$, les dates d'arrivées successives des utilisateurs dans le système ;
- $\Delta_j = T_j - T_{j-1}$ la durée inter-arrivée ;
- Services : $S_j^{(i)}$ désigne la durée mise par le serveur i pour s'occuper de l'utilisateur j ;

On a donc un flux d'arrivée, puis une file d'attente, puis une zone de service avec les différents serveurs et bien sûr un flux de départ. Lorsque tous les serveurs sont occupés, un nouvel utilisateur prend place dans la file d'attente.
Discipline de service : *FIFO* (First In, First Out). Attention à ne pas traduire "First In, First Out" par "Premier arrivé, Premier sortie" qui pourrait être faux mais plutôt par "Premier arrivé, Premier servi".
On utilise dans la suite la notation de Kendall qui synthétise le fonctionnement d'un tel système :

$$A/B/s/m$$

- A : Loi du flux d'entrée ;
- B : Loi du flux de départ ;
- s : Nombre de serveurs ;
- m : Capacité d'accueil limite, que l'on omet souvent.

$$A, B \in \{M, D, G\}$$

On note M pour Markov, D pour déterministe, G pour général.

- $A = M \iff$ Flux d'entrée est donné par un processus de Poisson d'intensité λ (flux d'entrée) ;
- $B = M \iff$ Durée de service indépendante et exponentielle (sans mémoire) de paramètre μ.

Dans la suite, on considère les files d'attente

$$M/M/*$$

avec $* = 1, \infty, s$.

On pose aussi $\rho = \dfrac{\lambda}{\mu}$ l'intensité du trafic. Pour analyser ces files d'attente, on introduit les quantités suivantes, en se plaçant dans le régime stationnaire de ces files ($t \to +\infty$, X_∞ : Nombre d'utilisateur dans le système est régime stationnaire).

- $L = \mathbb{E}(X_\infty)$ le nombre moyen d'utilisateur dans le système ;
- $L_q = \mathbb{E}(\max(X_\infty - s, 0))$ le nombre moyen d'utilisateur dans la file d'attente ;
- T le temps passé dans le système pour un utilisateur ;
- T_q le temps passé dans la file d'attente pour un utilisateur ;
- $W = \mathbb{E}(T)$;
- $W_q = \mathbb{E}(T_q)$.

Sous des conditions assez générales, on a

Théorème 7.5.1 (Formules de Little)**.** *En notant λ_e le flux d'entrée effectif lorsqu'il y a une restriction d'accueil (en général il n'y en aura pas et on aura $\lambda_e = \lambda$)*

$$L = \lambda_e W$$

$$L_q = \lambda_e W_q$$

$$W = W_q + \frac{1}{\mu}$$

$$L = L_q + \frac{\lambda_e}{\mu}$$

Démonstration : On peut déjà voir que la troisième formule est immédiate car le temps moyen d'une loi exponentielle de paramètre μ est $\frac{1}{\mu}$. Puis, si on admet un moment la 1ère et la 2ème, alors la 3ème coule de source.
Maintenant lorsqu'un utilisateur u arrive dans le système et qu'il y séjourne une durée T_u, en sortant il est arrivé N_{T_u} utilisateurs dans le système, où (N_t) est le processus de Poisson d'intensité λ qui alimente le système. Alors le nombre moyen d'utilisateurs dans le système est

$$L = \mathbb{E}(N_{T_u}) = \mathbb{E}\big(\mathbb{E}(N_{T_u}|T_u)\big) = \mathbb{E}(\lambda T_u) = \lambda W$$

File M/M/1 :

Le processus (X_t) est un processus de Markov pour lequel, $X_t = 0$ signifie serveur inoccupé, file d'attente vide et $X_t = x$ signifie serveur occupé et $x-1$ utilisateurs dans la file d'attente.

On reste dans cette situation une durée

$$\min(T^+, T^-) \sim \mathcal{E}(\lambda + \mu)$$

avec $T^+ \sim \mathcal{E}(\lambda)$ la date d'arrivée d'un nouvel utilisateur, $T^- \sim \mathcal{E}(\mu)$ la date de départ de l'utilisateur en service. On a $T^+ \perp\!\!\!\perp T^-$. Après cette durée $\min(T^+, T^-)$, (X_t) transite en $x+1$ avec probabilité $\frac{\lambda}{\lambda+\mu}$ ou en $x-1$ avec probabilité $\frac{\mu}{\lambda+\mu}$.

$\Longrightarrow$ (X_t) est un Processus de Naissance et de Mort sur $\mathbb{N}$ avec taux de naissance λ et taux de mort μ

$$\lambda_x = \lambda,\ \forall x \geq 0 \text{ et } \mu_x = \mu,\ \forall x \geq 1$$

Maintenant, pour déterminer la nature du processus (X_t), il suffit de préciser les natures des séries :

$$\sum_{x\geq 0} \prod_{i=1}^{x} \frac{\lambda_{i-1}}{\mu_i} = \sum_{x\geq 0} \left(\frac{\lambda}{\mu}\right)^x$$

$$\sum_{x\geq 0} \frac{\mu_1 \ldots \mu_{x-1}}{\lambda_1 \ldots \lambda_{x-1}} = \sum_{x\geq 0} \left(\frac{\mu}{\lambda}\right)^x$$

Si $\lambda < \mu$ alors le processus est récurrent positif et tout se passe bien. On aura $\rho = \frac{\lambda}{\mu} < 1$ et c'est l'hypothèse qu'on prendra dans la suite.

7.6 Théorie du renouvellement

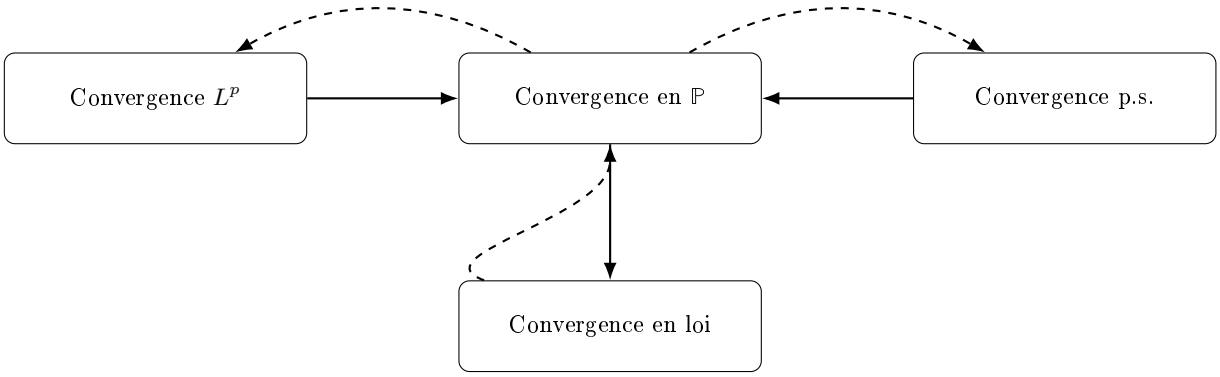
8 Processus Stochastique (M2)

Sources : **[1]** Processus et Calcul Stochastique, Jean-Christophe Breton, Université de Rennes - ENS Rennes, 2019-2020.
[2] Processus Stochastiques, Djalil Chafaï, Ecole Normale Supérieure, 2025-2026.

8.1 Rappels gaussiens

8.1.1 Rappels sur les convergences de variables aléatoires

- $X_n \to^{p.s.} X$ ssi $\mathbb{P}(\{\omega \in \Omega \mid X_n(\omega) \to X(\omega)\}) = 1$.
- $X_n \to^{\mathbb{P}} X$ ssi $\forall \epsilon > 0,\ \mathbb{P}(|X_n - X| > \epsilon) \to 0$.
- $X_n \to^{L^p} X$ ssi $\mathbb{E}(|X_n - X|^p) \to 0$.
- $X_n \to^{\mathcal{L}} X$ ssi $\forall f \in \mathcal{C}_b,\ \mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$ ssi $\forall t \in \mathbb{R},\ \varphi_{X_n}(t) \to \varphi_X(t)$ ssi $\forall t$ où F_X est continue, $F_{X_n}(t) \to F_X(t)$.



8.1.2 Variables gaussiennes

Définition 8.1.1. Une variable aléatoire X suit une loi normale standard $\mathcal{N}(0,1)$ si elle admet pour densité

$$t \in \mathbb{R} \ \mapsto \ \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right)$$

De façon générale, X suit la loi normale $\mathcal{N}(m, \sigma^2)$ $(m \in \mathbb{R})$ si elle admet pour densité

$$t \in \mathbb{R} \ \mapsto \ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-m)^2}{2\sigma^2}\right)$$

On note que si $\sigma^2 = 0$, la loi est dégénérée, $X \equiv m$. Sa loi est une mesure de Dirac en m : $\mathbb{P}_X = \delta_m$.

En effet, si $X \sim \mathcal{N}(m, \sigma^2)$, alors sa densité de probabilité est la fonction qui décrit la forme de la distribution normale centrée en m avec une "dispersion" mesurée par σ^2. Si bien que si on

prend l'asymptotique $[\sigma^2 \to 0]$ alors le dénominateur devient nul et on a une exponentielle de type $\exp(-\infty)$, SAUF que $t = m$ où elle vaut 1.
Intuitivement, le pic de la courbe devient de plus en plus étroit et haut, la probabilité se concentre de plus en plus autour de m. On a donc

$$X \sim \mathcal{N}(m, 0) \Rightarrow \mathbb{P}(X = m) = 1$$

ou encore, pour $X \sim \mathcal{N}(m, \sigma^2)$ avec $\sigma^2 = 0$

$$X = m \text{ presque sûrement.}$$

i.e. X est constante égale à m car l'écart-type nul signifie qu'il y a absence totale de dispersion autour de m.

Avant de passer à la suite, rappelons la formule de changement de variables pour les densités. Soit X une variable aléatoire de densité $f_X(x)$ et $Y = g(X)$, où g est bijective, strictement monotone et dérivable. On cherche la densité de Y.

$$\mathbb{P}(Y \leq y) = \mathbb{P}(X \leq g^{-1}(y)) = \int_{-\infty}^{g^{-1}(y)} f_X(x)dx$$

D'où, en supposant que f_X soit dérivable,

$$f_Y(y) = \frac{d}{dy}\left[\int_{-\infty}^{g^{-1}(y)} f_X(x)dx\right] = f_X(g^{-1}(y)).\frac{d}{dy}g^{-1}(y)$$

Si bien que, comme une densité doit être positive :

$$f_Y(y) = f_X(g^{-1}(y)).\left|\frac{d}{dy}g^{-1}(y)\right|$$

Proposition 8.1.1. *Si $X \sim \mathcal{N}(m, \sigma^2)$, alors on peut voir X comme la translatée et la dilatée de $X_0 \sim \mathcal{N}(0, 1)$ par $X = m + \sigma X_0$.*

Démonstration : Soit $X_0 \sim \mathcal{N}(0, 1)$, alors sa densité est

$$f_{X_0}(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

Posons $X = m + \sigma X_0$ et calculons f_X. Pour cela, on va utiliser la formule trouvée précédemment avec

$$X_0 = \frac{X - m}{\sigma}$$

On a

$$f_X(x) = f_{X_0}\left(\frac{x-m}{\sigma}\right).\left|\frac{d}{dx}\left(\frac{x-m}{\sigma}\right)\right| = \frac{1}{\sigma}f_{X_0}\left(\frac{x-m}{\sigma}\right)$$

D'où

$$f_X(x) = \frac{1}{\sigma}.\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right)$$

On retrouve bien la densité de la loi normale. □

Proposition 8.1.2. *Soit $X \sim \mathcal{N}(m, \sigma^2)$, alors :*

- $\mathbb{E}(X) = m$;
- $Var(X) = \sigma^2$;
- $\varphi_X(t) = \exp\left(imt - \dfrac{\sigma^2 t^2}{2}\right)$.

Démonstration : Pour l'espérance, on va centré réduire :

$$\mathbb{E}(X) = \int x f_X(x) dx$$

On pose $x = m + \sigma z$ et donc $dx = \sigma dz$. Si bien que

$$\mathbb{E}(X) = \int (m+\sigma z) f_X(\sigma z + m)\sigma dz = \int (m+\sigma z)\frac{1}{\sigma}.\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{(\sigma z + m) - m}{\sigma}\right)^2\right)\sigma dz$$

$$= \int (m+\sigma z)\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)dz = m \underbrace{\int \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)dz}_{=1,\text{ densité standard}} + \sigma \underbrace{\int \frac{1}{\sqrt{2\pi}} z \exp\left(-\frac{z^2}{2}\right)dz}_{=0,\text{ fonction impaire}}$$

Si bien que,

$$\mathbb{E}(X) = m$$

Pour la variance, on utilise la formule de Keonig-Huygens, ce qui nous laisse le moment d'ordre 2 à calculer et ce avec le même changement de variable :

$$\mathbb{E}(X^2) = \int x^2 f_X(x) dx = \int (m + \sigma z)^2 f_X(\sigma z + m)\sigma dz$$

En séparant les intégrales via le carré, on trouve

$$\mathbb{E}(X^2) = \sigma^2 + m^2$$

Et donc

$$\mathrm{Var}(X) = \sigma^2$$

Maintenant, concernant la fonction caractéristique

$$\varphi_X(t) = \mathbb{E}(e^{itX}) = \int e^{itx} f_X(x) dx$$

Par le même changement de variables

$$\varphi_X(t) = \int e^{it(\sigma z + m)}.\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)dz = e^{itm} \int \frac{1}{\sqrt{2\pi}} \exp\left(it\sigma z - \frac{z^2}{2}\right)dz$$

Puis, on complète le carré

$$-\frac{z^2}{2} + it\sigma z = -\frac{1}{2}\left(z^2 - 2it\sigma z\right) = -\frac{1}{2}\left(z^2 - 2it\sigma z + (it\sigma)^2 - (it\sigma)^2\right)$$

$$-\frac{z^2}{2} + it\sigma z = -\frac{1}{2}(z - it\sigma)^2 + \frac{1}{2}(it\sigma)^2 = -\frac{1}{2}(z - it\sigma)^2 - \frac{1}{2}(t\sigma)^2$$

Si bien que

$$\varphi_X(t) = e^{itm}.e^{-\frac{1}{2}t^2\sigma^2} \underbrace{\int \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(z - it\sigma)^2\right) dz}_{=1, \text{ translation densité normale}}$$

□

Notons qu'il y a une façon plus élégante de démontrer la formule de la fonction caractéristique de la loi normale. Il suffit d'identifier les fonctions holomorphes $\mathbb{E}(e^{zX})$ et $\exp\left(\frac{z^2}{2}\right)$ pour $z \in \mathbb{R}$ et considérer $z = ix$.

En effet, pour une variable aléatoire X, on peut considérer sa fonction génératrice des moments (MGF) définie pour tout réel $z \in \mathbb{R}$ (et même $z \in \mathbb{C}$ quand c'est possible i.e. quand l'espérance existe) par

$$M_X(z) = \mathbb{E}(e^{zX})$$

Dans le cadre d'une variable aléatoire quelconque, cette fonction n'est holomorphe que dans une bande autour de 0 :

$$D = \{z \in \mathbb{C} \mid |\mathrm{Re}(z)| < a\}$$

C'est-à-dire que la régularité analytique de la MGF dépend de la loi considérée et elle n'est donc pas toujours entière. Par exemple une variable aléatoire suivant une loi exponentielle. On a

$$f(x) = \lambda e^{-\lambda x} \mathbb{1}_{x \geq 0}$$

La MGF s'exprime

$$M_X(z) = \mathbb{E}(e^{zX}) = \int_0^\infty e^{zx}.\lambda e^{-\lambda x} dx = \lambda \int_0^\infty e^{x(z-\lambda)} dx$$

Cette intégrale convergence si et seulement si $\mathrm{Re}(z) < \lambda$. Dans ce cas, la MGF n'est holomorphe que sur le demi-plan $\mathrm{Re}(z) < \lambda$ et elle n'est donc pas entière.

Cependant, si la variable aléatoire X suit la loi normale $\mathcal{N}(m, \sigma^2)$ alors on peut faire une observation générale et une spécifique. L'observation générale est que la fonction caractéristique n'est rien d'autre que la MGF évaluée en $z = it \in i\mathbb{R}$:

$$\varphi_X(t) = \mathbb{E}(e^{itX}) = M_X(it)$$

L'autre "observation" revient à calculer explicitement la MGF pour $X \sim \mathcal{N}(m, \sigma^2)$.

$$M_X(z) = \mathbb{E}(e^{zX}) = \int e^{zx} f_X(x) dx$$

D'où,

$$M_X(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \int \exp\left(zx - \frac{(x-m)^2}{2\sigma^2}\right) dx$$

Par des calculs fastidieux, on développe le terme dans l'exponentielle (compléter le carré) que l'on reinjecte dans l'intégrale pour avoir cette hideuse expression :

$$M_X(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{1}{2}\sigma^2\left(z+\frac{m}{\sigma^2}\right) - \frac{m^2}{2\sigma^2}\right) \underbrace{\int \exp\left(-\frac{1}{2\sigma^2}\left[x-\sigma^2\left(z+\frac{m}{\sigma^2}\right)\right]^2\right)dx}_{=\sqrt{2\pi\sigma^2},\text{ densité d'une normale}}$$

Si bien que

$$M_X(z) = \exp\left(\frac{1}{2}\sigma^2\left(z+\frac{m}{\sigma^2}\right) - \frac{m^2}{2\sigma^2}\right) = \exp\left(\frac{1}{2}\sigma^2 z^2 + mz\right), \text{ pour } z \in \mathbb{C}.$$

Puis en faisant $z = it$ et en utilisant le fait rappeler plus haut, on obtient

$$\varphi_X(t) = \exp\left(imt - \frac{\sigma^2 t^2}{2}\right)$$

Pour conclure cet aparté sur la MGF, on rappelle que la queue de distribution désigne le comportement de la loi de probabilité pour des valeurs extrêmes. Plus précisément, la queue à droite correspond à $\mathbb{P}(X > x)$ quand $x \to \infty$ et la gauche à gauche correspond à $\mathbb{P}(X < -x)$ quand $x \to \infty$. Autrement dit, la queue mesure à quelle vitesse décroit la probabilité que X soit très loin de la moyenne. Quel rapport avec la MGF ?

La MGF est définie par

$$M_X(z) = \int_{\mathbb{R}} e^{zx} f_X(x)dx$$

Or, l'exponentielle e^{zx} croît excessivement vite quand $x \to \infty$, encore plus si $\mathrm{Re}(z) > 0$. Ainsi, pour que l'intégrale converge, il faut que la densité décroisse plus rapidement que l'exponentielle. Si la queue décroît lentement, l'intégrale peut diverger et donc la MGF n'existe pas.

Si bien que, ce qui compte vraiment pour l'existence de la MGF, ce n'est pas juste le support, mais la vitesse de décroissance de la queue de distribution.

Par exemple, si $X \sim \mathcal{C}(0,1)$ alors

$$f_X(x) = \frac{1}{\pi(1+x^2)} \sim_\infty \frac{1}{x^2}$$

ce qui est beaucoup trop lent (même l'espérance n'existe pas) et donc la MGF n'existe nulle part sauf en $z = 0$.

Théorème 8.1.1 (Existence de la MGF)**.** *Soit X une variable aléatoire réelle. Le domaine de définition de M_X est un intervalle ouvert convexe centré en 0 dans $\mathbb{R}$, et plus généralement une bande verticale ouverte convexe dans $\mathbb{C}$*

$$D = \{z \in \mathbb{C} \mid \mathbb{E}(e^{Re(z)X}) < \infty\}$$

Autrement dit,

$$M_X(z) = \int \exp(Re(z)x) d\mathbb{P}_X(x) \text{ existe } \iff \mathbb{E}(e^{Re(z)X}) < \infty$$

La MGF est bien définie et holomorphe sur ce domaine.

Il existe aussi des résultats plus fins concernant la MGF, notamment sur son lien avec les moments de la loi : Si tous les moments existent, cela n'impliquent pas qu'elle soit définie partout sur $\mathbb{R}$; Il existe des variables aléatoires telles que tous les moments existent mais telle que la MGF diverge pour tout $t \neq 0$, dû par exemple à une croissance trop rapide des moments qui empêche la série de Taylor de la MGF de converger hors de 0. On peut aussi citer le théorème de Hardy dans le contexte du problème des moments (si la MGF n'est définie que sur un intervalle alors elle ne détermine pas forcément la loi)...

Repassons à des résultats sur les variables gausiennes.

Proposition 8.1.3 (Moments de $\mathcal{N}(0,1)$)**.** *Soit $Z \sim \mathcal{N}(0,1)$. Alors,*

$$\mathbb{E}(Z^{2n}) = \frac{(2n)!}{2^n n!},\ \mathbb{E}(Z^{2n+1}) = 0$$

Démonstration : La densité est

$$f_Z(x) = \frac{1}{\sqrt{1\pi}} \exp\left(-\frac{x^2}{2}\right)$$

La densité est donc paire dans ce cas et Z^{2n+1} est impaire, donc comme l'intervalle d'intégration est symétrique

$$\mathbb{E}(Z^{2n+1}) = \int x^{2n+1} f_Z(x) dx = 0$$

Pour les moments pairs, raisonnons par récurrence sur n

$$\forall n \in \mathbb{N},\ H_n\ :\ "\mathbb{E}(Z^{2n}) = \frac{(2n)!}{2^n n!}"$$

Initialisation : C'est bon, H_0 est vraie.
Hérédité : Soit $n \in \mathbb{N}$ fixé tel que H_n soit vraie. Montrons H_{n+1}.
On suppose donc

$$\mathbb{E}(Z^{2n}) = \frac{(2n)!}{2^n n!}$$

On veut montrer que

$$\mathbb{E}(Z^{2n+2}) = \frac{(2n+2)!}{2^{n+1}(n+1)!} = \frac{(2n+2)(2n+1)}{2(n+1)} \mathbb{E}(Z^{2n}) = (2n+1)\mathbb{E}(Z^{2n})$$

Pour ça on fait une IPP

$$\mathbb{E}(Z^{2n}) = \frac{1}{\sqrt{2\pi}} \int x^{2n} \exp\left(-\frac{x^2}{2}\right) dx$$

$$\int x^{2n}\exp\left(-\frac{x^2}{2}\right)dx = \left[\frac{x^{2n+1}}{2n+1}\exp\left(-\frac{x^2}{2}\right)\right]_{-\infty}^{+\infty} + \frac{1}{2n+1}\int x^{2n+2}\exp\left(-\frac{x^2}{2}\right)dx$$

D'où, comme le crochet tend vers 0

$$(2n+1)\int x^{2n}\exp\left(-\frac{x^2}{2}\right)dx = \int x^{2n+2}\exp\left(-\frac{x^2}{2}\right)dx$$

Et donc, H_{n+1} est vraie et H_n est héréditaire.
□

Proposition 8.1.4 (Moments d'ordre n de $\mathcal{N}(m,\sigma^2)$)**.** *Soit $X \sim \mathcal{N}(m,\sigma^2)$,*

$$\mathbb{E}(X^n) = \sum_{n=0}^{\left\lfloor\frac{n}{2}\right\rfloor}\binom{n}{2k}m^{n-2k}\sigma^{2k}\frac{(2n)!}{2^n n!}$$

Démonstration : On raisonne par translation et dilation :

$$X = m + \sigma X,\ Z \sim \mathcal{N}(0,1)$$

Alors,

$$\mathbb{E}(X^n) = \mathbb{E}((m+\sigma X)^n)$$

Par la formule du binôme de Newton ainsi que par linéarité de l'espérance

$$\mathbb{E}(X^n) = \sum_{k=0}^{n}\binom{n}{k}m^{n-k}\sigma^k\mathbb{E}(Z^k)$$

On utilise ensuite la formule pour les moments de Z qui nous laisse uniquement les termes pairs

$$\mathbb{E}(X^n) = \sum_{k=0}^{\left\lfloor\frac{n}{2}\right\rfloor}\binom{n}{2k}m^{n-2k}\sigma^{2k}\mathbb{E}(Z^{2k})$$

□

On peut aussi montrer que pour $X \sim \mathcal{N}(0,\sigma^2)$ alors $\mathbb{E}(X^n) = \sigma^2\mathbb{E}(Z^n)$ si n est pair et 0 sinon, en posant simplement $X = \sigma Z$.

Maintenant, essayons d'approximer la queue normale. On a connaissance de 3 bornes possibles pour celle-ci.

Proposition 8.1.5 (Borne de Mill)**.** *Soit $X \sim \mathcal{N}(0,1)$ et $x > 0$. Alors,*

$$\mathbb{P}(X \geq x) = \int_x^\infty \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right)dt \leq \frac{1}{x}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$$

Démonstration : Les fonctions considérées sont effectivement $\mathcal{C}^1$ donc on peut faire une IPP.

$$\int_x^\infty \frac{1}{t}.t\exp\Big(-\frac{t^2}{2}\Big)dt = \Big[\frac{1}{t}.\Big(-\exp\Big(-\frac{t^2}{2}\Big)\Big)\Big]_x^\infty - \int_x^\infty \frac{1}{t^2}\exp\Big(-\frac{t^2}{2}\Big)dt$$
$$= \frac{1}{x}\exp\Big(-\frac{x^2}{2}\Big) - \int_x^\infty \frac{1}{t^2}\exp\Big(-\frac{t^2}{2}\Big)dt$$

Si bien que

$$\mathbb{P}(X \geq x) = \frac{1}{\sqrt{2\pi}}\Big(\frac{1}{x}\exp\Big(-\frac{x^2}{2}\Big) - \int_x^\infty \frac{1}{t^2}\exp\Big(-\frac{t^2}{2}\Big)dt\Big)$$

D'où, comme l'intégrale est positive,

$$\mathbb{P}(X \geq x) < \frac{1}{\sqrt{2\pi}}\frac{1}{x}\exp\Big(-\frac{x^2}{2}\Big)$$

□

Maintenant, on a du raffinage pour $x \geq 1$.

Proposition 8.1.6 (Borne de Mill améliorée)**.** *Maintenant, prenons $x \geq 1$ alors*

$$\mathbb{P}(X \geq x) \leq \Big(\frac{1}{x} - \frac{1}{x^2}\Big)\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{x^2}{2}\Big)$$

Démonstration : On repart de

$$\mathbb{P}(X \geq x) = \frac{1}{\sqrt{2\pi}}\Big(\frac{1}{x}\exp\Big(-\frac{x^2}{2}\Big) - \int_x^\infty \frac{1}{t^2}\exp\Big(-\frac{t^2}{2}\Big)dt\Big)$$

Aussi valable dans le cadre où $x \geq 1$. Par inversion il vient que

$$\forall t \geq x,\ \frac{1}{t^2} \leq \frac{1}{x^2}$$

Si bien que,

$$\int_x^\infty \frac{1}{t^2}\exp\Big(-\frac{t^2}{2}\Big)dt \leq \frac{1}{x^2}\int_x^\infty \exp\Big(-\frac{t^2}{2}\Big)dt = \frac{1}{x^2}\sqrt{2\pi}\mathbb{P}(X \geq x)$$

En injectant, il vient que

$$\mathbb{P}(X \geq x) \leq \frac{1}{\sqrt{2\pi}}\frac{1}{x}\exp\Big(-\frac{x^2}{2}\Big) - \frac{1}{x^2}\mathbb{P}(X \geq x)$$

Ainsi,

$$\mathbb{P}(X \geq x)\Big(1 + \frac{1}{x^2}\Big) \leq \frac{1}{\sqrt{2\pi}}\frac{1}{x}\exp\Big(-\frac{x^2}{2}\Big)$$

Puis,

$$\frac{1}{1+\frac{1}{x^2}} = 1 - \frac{1}{x^2} + o\Big(\frac{1}{x^2}\Big)$$

D'où l'inégalité.□

Ce résultat est enfait une version modifiée du developpement asymptotique de la queue normale. En effet, en faisant des intégrations par parties successives, on obtient

$$\mathbb{P}(X \geq x) \sim_{x\to\infty} \frac{\exp\left(-\frac{x^2}{2}\right)}{x\sqrt{2\pi}} \sum_{k=0}^{\infty} \frac{(-1)^k}{x^{2k}} \frac{(2k)!}{2^k k!}$$

On a aussi une autre borne que l'on déduit de la première, simplifiée mais très utile :

Proposition 8.1.7. *Toujours dans le cadre où $x \geq 1$, on a l'inégalité suivante*

$$\mathbb{P}(X \geq x) \geq \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

Passons maintenant à d'autres résultats et opérations.

Proposition 8.1.8. *Soient $N_1 \sim \mathcal{N}(m_1, \sigma_1^2)$ et $N_2 \sim \mathcal{N}(m_2, \sigma_2^2)$ indépendantes. Alors*

$$N_1 + N_2 \sim \mathcal{N}(m_1 + m_2, \sigma_1^2 + \sigma_2^2)$$

Démonstration : Via les fonctions caractéristiques et l'indépendance, on a

$$\varphi_{N_1+N_2}(t) = \varphi_{N_1}(t)\varphi_{N_2}(t) = \varphi_{\mathcal{N}(m_1+m_2,\sigma_1^2+\sigma_2^2)}(t)$$

□

Proposition 8.1.9. *Soit (X_n) une suite de variables aléatoires normales de lui $\mathcal{N}(m_n, \sigma_n^2)$.*

1. *La suite (X_n) converge en loi si et seulement si $m_n \to m \in \mathbb{R}$ et $\sigma_n^2 \to \sigma^2 \in \mathbb{R}^+$. La loi limite est alors $\mathcal{N}(m, \sigma^2)$.*
2. *Si la suite converge en probabilité vers X, la convergence a lieu dans tous les espaces L^p, $p < \infty$.*

Démonstration : On montre d'abord le premier résultat par double-implication.
On rappelle que le théorème de Lévy établit que la convergence en loi $X_n \Longrightarrow X$ est équivalente à avoir pour tout $t \in \mathbb{R}$

$$\varphi_{X_n}(t) = \exp\left(im_n t - \frac{\sigma_n^2}{2}t^2\right) \to \varphi_X(t)$$

A partir de ce résultat, on veut pouvoir isoler les paramètres m_n et σ_n^2. Pour cela, on passe au module. Comme φ_X est continue et $\varphi_X(0) = 1$, il existe $t \neq 0$ tel que $|\varphi_X(t)| \neq 0$. Pour ce t, on a alors

$$|\varphi_{X_n}(t)| = \left|\exp(im_n t)\exp\left(-\frac{\sigma_n^2}{2}t^2\right)\right| = \exp\left(-\frac{\sigma_n^2}{2}t^2\right) \to |\varphi_X(t)|$$

Ainsi, en passant au ln, on en déduit que la limite suivante existe

$$\lim_{n\to\infty} \sigma_n^2 = -\frac{2}{t^2}\ln|\varphi_X(t)| := \sigma^2, \text{ existe}$$

Par suite, on a aussi

$$\frac{\varphi_{X_n}(t)}{|\varphi_{X_n}(t)|} = \exp(im_n t) \to \exp\left(\frac{\sigma^2}{2}t^2\right)$$

La convergence de $\varphi_{X_n}(t)$ implique que la quantité doit aussi converger. Cependant cela n'est possible que si m_n est bornée.
Raisonnons par l'absurde en supposant que (m_n) ne soit pas bornée. On construit alors une sous-suite telle que $m_{n_k} \geq k$ et $m_{n_k} \to \infty$.
Pour une loi normale centrée en m_n, la médiane est m_n. Donc,

$$\mathbb{P}(X_{n_k} \geq m_{n_k}) = \frac{1}{2}$$

Si bien que, si $\eta > 0$,

$$\mathbb{P}(X_{n_k} \geq \eta) \geq \mathbb{P}(X_{n_k} \geq m_{n_k}) = \frac{1}{2}$$

Mais comme $m_{n_k} \to \infty$, il existe un rang à partir duquel $m_{n_k} \geq \eta$. Donc,

$$\liminf_{k\to\infty} \mathbb{P}(X_{n_k} \geq \eta) \geq \frac{1}{2},\ \forall \eta > 0$$

Cela implique que, pour toute valeur de $\eta > 0$, la probabilité que X_{n_k} soit plus grande que η est au moins $\frac{1}{2}$ dans la limite. Intuitivement, la masse de probabilité s'échappe à droite ves ∞.
Cependant, si X_{n_k} converge en loi vers X pour une variable réelle, alors

$$\mathbb{P}(X \geq \eta) = \lim \mathbb{P}(X_{n_k} \geq \eta)$$

Si bien qu'on aurait $\mathbb{P}(X \geq \eta) \geq \frac{1}{2}$ et ce, pour tout $\eta > 0$. Cependant, c'est impossible pour une variable aléatoire réelle : la probabilité qu'elle soit plus grande que tous les $\eta > 0$ ne peut pas être supérieure à $\frac{1}{2}$ à chaque fois. Cela impliquerait qu'elle soit infinie presque sûrement. C'est-à-dire,

$$\forall \eta > 0,\ \mathbb{P}(X \geq \eta) \geq \frac{1}{2} \Rightarrow \mathbb{P}(X = +\infty) \geq \frac{1}{2}$$

Ce qui est impossible. Donc la suite (X_n) ne peut pas converger en loi vers une variable réelle, ce qui est absurde car $\mathbb{P}(X \in \mathbb{R}) = \lim \mathbb{P}(X_n \in \mathbb{R}) = 1$.
Cela force la suite (m_n) à être bornée.
Dès lors, si m et m' sont deux valeurs d'adhérence de (m_n), en passant à la limite sur les bonnes sous-suites, on doit avoir pour tout $t \in \mathbb{R}$, $\exp(imt) = \exp(im't)$. Cela exige $m = m'$. Il y a donc unicité de la valeur d'adhérence, c'est-à-dire existence de la limite m de m_n.
Finalement, par passage à la limite,

$$\varphi_X(t) = \exp\left(imt - \frac{\sigma^2}{2}t^2\right)$$

Ce qui assure aussi la loi de X.
Le sens réciproque est immédiat.
Pour la deuxième propriété, on écrit $X_n = m_n + \sigma_n N_n$ avec $N_n \sim \mathcal{N}(0,1)$. Comme X_n converge en loi, les suites (m_n) et (σ_n) sont bornées. Par convexité pour $q > 1$

$$|\sigma_n N_n + m_n|^q = \left|\frac{1}{2}(2\sigma_n N_n + 2m_n)\right|^q \leq 2^{q-1}(|\sigma_n|^q |N_n|^q + |m_n|^q)$$

C'est une inégalité très utile à avoir en tête car elle est optimale en termes d'ordre de grandeur (l'exposant ne peut pas être amélioré uniformément sur $\mathbb{R}^2$). Elle est très utilisée pour prouver des majorations de moments.
Maintenant, l'expression des moments de N_n assure que

$$\sup_n \mathbb{E}(|X_n|^q) < \infty, \ \forall q \geq 1$$

Comme la convergence en probabilité donne la convergence presque sûrement d'une sous-suite X_{n_k}, par le lemme de Fatou

$$\mathbb{E}(|X|^q) = \mathbb{E}(\lim |X_{n_k}|^q) \leq \liminf \mathbb{E}(|X_{n_k}|^q) \leq \sup \mathbb{E}(|X_{n_k}|^q) \leq \sup \mathbb{E}(|X_n|^q) < \infty$$

Soit $p \geq 1$, la suite $|X_n - X|^p$ converge vers 0 en probabilité et est uniformémement intégrable car bornée dans L^2 ($q = 2p$). Elle converge donc dans L^1 vers 0.
□

Théorème 8.1.2 (Loi faible des grands nombres). *Soit (X_n) une suite de variables aléatoires indépendantes et identiquement distribuées intégrables, i.e. $\mathbb{E}(|X_1|) < \infty$, alors*

$$\frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \to^{\mathbb{P}} \mathbb{E}(X_1)$$

Théorème 8.1.3 (Loi forte des grands nombres). *Soit (X_n) une suite de variables aléatoire indépendantes et identiquement distribuées intégrables, i.e. $\mathbb{E}(|X_1|) < \infty$, alors*

$$\frac{S_n}{n} = \frac{1}{n}\sum_{k=1}^{n} X_i \to^{p.s.,L^1} \mathbb{E}(X_1)$$

Le caractère universel de la loi normale est illustré par le résultat suivant. Il montre que la loi normale standard contrôle les fluctuations par rapport à leur moyenne des effets cumulés d'un phénomène aléatoire répété avec des répétitions indépendantes. Il exprime l'omniprésence de la loi normale en statistique (tests statistiques, intervalles de confiance...), Physique, Economie...

Théorème 8.1.4 (Théorème Central Limite - Lindeberg-Lévy). *Soit (X_n) une suite de variables aléatoires iid, d'espérance m et de variance finie $\sigma^2 > 0$ (carré intégrable). Soit $S_n = X_1 + ... + X_n$. Alors*

$$\frac{S_n - nm}{\sqrt{\sigma^2 n}} \Longrightarrow \mathcal{N}(0,1)$$

Il est à noter que l'on peut reécrire ce théorème de la manière suivante :

$$\sqrt{n}\left(\frac{S_n}{n} - \mathbb{E}(X_1)\right) \to^{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

de façon équivalente

$$\frac{\sqrt{n}}{\sigma}\left(\frac{S_n}{n} - \mathbb{E}(X_1)\right) \to^{\mathcal{L}} \mathcal{N}(0, 1)$$

ou encore,

$$\forall [a,b] \subseteq \mathbb{R},\ \mathbb{P}\left(\frac{\sqrt{n}}{\sigma}\left(\frac{S_n}{n} - \mathbb{E}(X_1)\right) \in [a,b]\right) \to \frac{1}{2\pi}\int_a^b \exp\left(-\frac{x^2}{2}\right)dx$$

$\to$ Pour voir une preuve de ces 3 derniers résultats, se référer au premier chapitre de ce PDF, plus particulièrement à la partie 1.8.

Le TCL complète la LGN : en effet, la LGN donne $\frac{S_n}{n} \to m$, i.e. $S_n - nm \sim 0$. Le TCL donne la vitesse de cette convergence (en loi), elle est en $\sqrt{n}$. Noter que la convergence est p.s. dans la LGN et en loi (donc plus faible) dans le TCL.
La loi $\mathcal{N}(0,1)$ apparaît à la limite dans le TCL alors que les variables aléatoires X_i sont de lois arbitraires (de carré intégrables) : ce résultat justifie donc le caractère universel de la loi normale comme dit ci-dessus. Ell modélise les petites variations de n'importe quelle loi (avec un moment d'ordre 2) par rapport à sa moyenne.

Le TCL nous donne alors la règle d'approximation suivante : La somme S_n d'une suite de variable aléatoire iid L^2 de moyenne m et de variance σ^2 s'approxime par

$$S_n \sim \mathcal{N}(nm, n\sigma^2)$$

On voit que le TCL de Lindeberg-Lévy requiert tout de même une condition forte qu'est la finitude de la variance. On peut faire mieux.

Théorème 8.1.5 (TCL avec condition de Lindeberg)**.** *Soit (X_n) une suite de variables aléatoires indépendantes telles que $\mathbb{E}(X_n) = \mu_n$ et $Var(X_n) = \sigma_n^2$. Posons*

$$S_n = \sum_{i=1}^n X_i \text{ et } s_n^2 = \sum_{i=1}^n \sigma_i^2$$

Si la condition de Lindeberg suivante est satisfaite

$$\forall \varepsilon > 0,\ \frac{1}{s_n^2}\sum_{i=1}^n \mathbb{E}((X_i - \mu_i)^2 \mathbb{1}_{|X_i - \mu_i| > \varepsilon s_n}) \to 0$$

Alors,

$$\frac{S_n}{s_n} \longrightarrow^{\mathcal{L}} \mathcal{N}(0,1)$$

La condition de Lindeberg contrôle l'impact des grandes déviations (les queues) dans chaque variable, i.e. la somme mesure l'impact des grandes valeurs de X_i (qui dépassent εs_n) sur la variance totale. Intuitivement, elle empêche qu'un petit nombre de termes dominateurs (à forte variance ou grosse amplitude) pertubent la convergence globale. En effet, si un seul X_i est trop lourd en probabilité alors il peut fausser la normalité asymptotique.

Voyons le rôle crucial de la condition de Lindeberg à travers un exemple et un contre-exemple.

Exemple 8.1 (Cas des variables iid et L^2). Tout d'abord, si les lois sont iid alors la condition de Lindeberg est automatiquement satisfaite : Si $X_i \sim X$ pour tout i alors avec $\mu = \mathbb{E}(X)$ et $\sigma^2 = \mathrm{Var}(X)$ il vient $s_n^2 = n\sigma^2$. La condition devient

$$\frac{1}{n\sigma^2}\sum_{i=1}^{n}\mathbb{E}((X_i-\mu)^2\mathbb{1}_{|X_i-\mu|>\varepsilon\sqrt{n\sigma^2}}) = \frac{1}{\sigma^2}\mathbb{E}((X-\mu)^2\mathbb{1}_{|X-\mu|>\varepsilon\sqrt{n\sigma^2}})$$

On note

$$\lambda_n = \varepsilon\sqrt{n\sigma^2} \to \infty$$

On a

$$(X-\mu)^2\mathbb{1}_{|X-\mu|>\lambda_n} \to 0 \text{ et dominé par } (X-\mu)^2 \in L^1$$

Par le théorème de convergence dominée

$$\mathbb{E}((X-\mu)^2\mathbb{1}_{|X-\mu|>\lambda_n}) \to 0$$

Et donc la condition de Lindeberg est satisfaite.

Exemple 8.2 (Contre-exemple). On veut une suite de variables aléatoires indépendantes, centrées telles que $\mathrm{Var}(X_n) < \infty$ mais la condition de Lindeberg échoue et la somme normalisée ne converge pas vers une gaussienne.
Prenons,

$$X_i = \begin{cases} \sqrt{i}, & \text{avec probabilité } \dfrac{1}{2i}, \\ -\sqrt{i}, & \text{avec probabilité } \dfrac{1}{2i}, \\ 0, & \text{avec probabilité } 1-\dfrac{1}{i}. \end{cases}$$

Alors, $\mathbb{E}(X_i) = 0$ donc les variables sont centrées, $\mathbb{E}(X_i^2) = i\frac{1}{i} = 1$ donc la variance est finie. Justifions qu'elles sont indépendantes.
Pour chaque $i \geq 1$ définis l'espace discret

$$\Omega_i = \{-\sqrt{i}, 0, \sqrt{i}\}$$

muni de la tribu $\mathcal{F}_i = 2^{\Omega_i}$ et de la mesure de probabilité $\mathbb{P}_i$ donnée par

$$\mathbb{P}_i(\{\sqrt{i}\}) = \frac{1}{2i},\ \mathbb{P}_i(\{-\sqrt{i}\}) = \frac{1}{2i},\ \mathbb{P}_i(\{0\}) = 1-\frac{1}{i}$$

Considérons l'espace produit

$$(\Omega, \mathcal{F}, \mathbb{P}) = \left(\prod_{i\geq 1}\Omega_i, \bigotimes_{i\geq 1}\mathcal{F}_i, \bigotimes_{i\geq 1}\mathbb{P}_i\right)$$

i.e. l'espace produit muni de la mesure produit (dont l'existence est garantie par le théorème de Kolmogorov pour produits dénombrables). Par construction, X_i prend exactement les valeurs et probabilités souhaitées et la mesure produit assure que les coordonnées sont indépendantes :

$$\forall i_1, ..., i_k \text{ distincts } \forall a_{i_k} \in \Omega_{i_k},\ \mathbb{P}(X_{i_1} = a_{i_1}, ..., X_{i_n} = a_{i_n}) = \prod_{k=1}^{n} \mathbb{P}(X_{i_k} = a_{i_k})$$

$\rightarrow$ En construisant les X_i comme coordonnées indépendantes sur un produit probabilisé, on obtient une suite indépendante (via factorisation des probabilités finies).
Maintenant, les X_i ne sont pas identiquement distribuées, car les queues changent à chaque rang. La variance totale est $s_n^2 = n$. La variable normalisée est

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$$

Fixons $\varepsilon > 0$. Alors

$$\mathbb{1}_{|X_i| > \varepsilon\sqrt{n}} = \begin{cases} 1, & \text{si } \sqrt{i} > \varepsilon\sqrt{n} \Leftrightarrow i > \varepsilon^2 n, \\ 0, & \text{sinon} \end{cases}$$

Donc,

$$\frac{1}{s_n^2} \sum_{i=1}^{n} \mathbb{E}(X_i^2 . \mathbb{1}_{|X_i| > \varepsilon s_n}) = \frac{1}{n} \sum_{i > \varepsilon^2 n} \mathbb{E}(X_i^2) = \frac{1}{n} \sum_{i > \varepsilon^2 n} 1 = \frac{n - \lfloor \varepsilon^2 n \rfloor}{n} \to 1 - \varepsilon^2$$

Ce terme ne tend pas vers 0 donc la condition de Linderberg n'est pas satisfaite.
Z_n ne suit alors pas de TCL.

Il existe aussi une version du TCL pour les variables aléatoires non-identiquement distribuées, avec une condition plus simple à vérifier, dîtes de Lyapunov.

Théorème 8.1.6 (TCL avec condition de Lyapunov)**.** *Soit (X_n) une suite de variables aléatoires indépendantes telles que,*

$$\forall i,\ \mu_i = \mathbb{E}(X_i),\ \sigma_i^2 = Var(X_i) < \infty$$

Posons,

$$\forall n \geq 1,\ S_n = \sum_{i=1}^{n} (X_i - \mu_i),\ s_n^2 = Var(S_n) = \sum_{i=1}^{n} \sigma_i^2$$

Supposons que $s_n^2 \to \infty$ et qu'il existe $\delta > 0$ tel que la condition de Lyapunov suivante soit satisfaite :

$$\frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}\Big(|X_i - \mu_i|^{2+\delta} \Big) \to 0$$

Alors, la somme normalisée converge en loi vers la loi normale standard

$$\frac{S_n}{s_n} \to^{\mathcal{L}} \mathcal{N}(0, 1)$$

On peut montrer que cette condition implique celle de Lindeberg. En effet, plaçons nous dans le cadre décrit par les 2 théorème, en supposant que $s_n^2 \to \infty$ et la condition de Lyapunov :

$$\exists \delta > 0 \mid \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}\Big(|X_i - \mu_i|^{2+\delta} \Big) \to 0$$

Fixons alors $\varepsilon > 0$. Pour chaque i, on a pour tout $u > 0$ et tout $\delta > 0$

$$\mathbb{1}_{\{u > \varepsilon s_n\}} \leq \frac{u^\delta}{(\varepsilon s_n)^\delta}$$

Prenons alors $u = |X_i - \mu_i|$ et multiplions par $(X_i - \mu_i)^2$

$$(X_i - \mu_i)^2 \mathbb{1}_{\{|X_i - \mu_i| > \varepsilon s_n\}} \leq \frac{|X_i - \mu_i|^{2+\delta}}{(\varepsilon s_n)^\delta}$$

D'où,

$$\mathbb{E}\big((X_i - \mu_i)^2 \mathbb{1}_{\{|X_i - \mu_i| > \varepsilon s_n\}} \big) \leq \frac{\mathbb{E}\big(|X_i - \mu_i|^{2+\delta} \big)}{(\varepsilon s_n)^\delta}$$

Puis,

$$\frac{1}{s_n^2} \sum_{i=1}^{n} \mathbb{E}\big((X_i - \mu_i)^2 \mathbb{1}_{\{|X_i - \mu_i| > \varepsilon s_n\}} \big) \leq \underbrace{\frac{1}{(\varepsilon s_n)^\delta} \frac{1}{s_n^2} \sum_{i=1}^{n} \mathbb{E}\big(|X_i - \mu_i|^{2+\delta} \big)}_{\to 0, \text{ par Lyapunov}}$$

Si bien que, par théorème d'encadrement

$$\frac{1}{s_n^2} \sum_{i=1}^{n} \mathbb{E}\big((X_i - \mu_i)^2 \mathbb{1}_{\{|X_i - \mu_i| > \varepsilon s_n\}} \big) \to 0$$

Ce qui est exactement la condition de Lindeberg.

On peut construire une variable aléatoire qui vérifie la condition de Lindeberg mais pas celle de Lyapunov. Elle est assez difficile à construire donc on la passe.

Il est à noter que ces théorèmes peuvent être énoncé dans un cadre plus général de tableau triangulaire de variables aléatoires $(X_{n,k})$ indépendantes par ligne (pour chaque n, les $X_{n,1}, ..., X_{n,r_n}$ sont indépendantes).
Il y a aussi d'autres TCL que l'on pourrait énoncer, par exemple un TCL pour les fonctionnelle. On ne le fera pas ici.

On donne, pour finir, une application canonique classique du Théorème Central Limite dans le cadre des suites binomiales.

Proposition 8.1.10 (Moivre-Laplace)**.** *Soit $S_n \sim \mathcal{B}(n,p)$, i.e. $S_n = X_1 + ... + X_n$ où les X_i sont indépendantes et suivent $\mathcal{B}(p)$ avec $p \in (0,1)$ fixé. On note $\mu_n = \mathbb{E}(S_n) = np$ et $\sigma_n^2 = Var(S_n) = np(1-p)$.*
Alors,

$$\forall a, b \in \mathbb{R}, \ \lim_{n \to \infty} \mathbb{P}\left(a \leq \frac{S_n - \mu_n}{\sigma_n} \leq b \right) = F_{\mathcal{N}(0,1)}(b) - F_{\mathcal{N}(0,1)}(a)$$

où $F_{\mathcal{N}(0,1)}$ est la fonction de répartition de la loi normale standard.

Autrement dit,

$$\frac{S_n - np}{\sqrt{np(1-p)}} \to^{\mathcal{L}} \mathcal{N}(0,1)$$

Ce qui est exactement le théorème central limite appliqué à la somme des $\mathcal{B}(p)$.
En appliquant une approximation de Stirling sur les coefficients binomiaux, on peut aussi obtenir une forme locale de Moivre-Laplace qui est plus "fine". Cela nous permet aussi de faire des application numérique avec une excellente précision.

8.1.3 Vecteurs gaussiens

On considère un vecteur aléatoire dans $\mathbb{R}^n$. Lorsqu'on le muni de son produit scalaire canonique, $\mathbb{R}^n$ est un espace euclidien.
Pour $a, b \in \mathbb{R}^n$ on note leur produit scalaire comme suit

$$\langle a, b \rangle = \sum_{i=1}^{n} a_i b_i$$

On peut généraliser cette section à un espace E euclidien (si $\dim(E) = n$ alors $E \sim \mathbb{R}^n$).

Définition 8.1.2 (Vecteur gaussien)**.** Un vecteur aléatoire $X = (X_1, ..., X_d)$ est gaussien (de dimension d) si et seulement si toutes les combinaisons linéaires de ses coordonnées

$$\forall a \in \mathbb{R}^d,\ \langle a, X \rangle = a^T X = a_1 X_1 + ... + a_d X_d$$

suivent une loi gaussienne $\mathcal{N}(m_a, \sigma_a^2)$ dans $\mathbb{R}$.
Dans un cadre euclidien, X vecteur à valeurs dans E est gaussien si et seulement si pour tout $a \in E$, $\langle a, X \rangle$ suit une loi guassienne.

En particulier, chaque marginale X_i suit une loi normale et a donc un moment d'ordre 2 fini. Les moments joints $\mathbb{E}(X_i X_j)$, pour $1 \leq i, j \leq n$ sont donc bien définis (par Cauchy-Schwarz) et on peut donc définir la matrice de covariance.

Définition 8.1.3 (Matrice de covariance)**.** La matrice de covariance d'un vecteur gaussien de dimension d est la matrice carrée symétrique, semi-définie positive

$$K = \big(\mathrm{Cov}(X_i, X_j)\big)_{1 \leq i,j \leq n}$$

Si $\det(K) = 0$ alors le vecteur est dit dégénéré.
L'espérance de X est le vecteur des espérances de ses marginales

$$m = \mathbb{E}(X) = \big(\mathbb{E}(X_1), ..., \mathbb{E}(X_n)\big)$$

Si $m = 0$ alors on dit que le vecteur est centré.
On a biensûr la définition plus explicite suivante :

$$K = \mathrm{Cov}(X) = \mathbb{E}\big((X - m)(X - m)^T\big)$$

$$K_{ij} = \mathrm{Cov}(X_i, X_j) = \mathbb{E}\big((X_i - m_i)(X_j - m_j)\big)$$

On rappelle qu'une matrice carrée A est symétrique si et seulement si $A = A^T$. C'est-à-dire, pour tout i, j, $A_{ij} = A_{ji}$. C'est toujours vérifié pour la matrice de covariance car

$$K_{ij} = \mathbb{E}\big((X_i - m_i)(X_j - m_j)\big) = \mathbb{E}\big((X_j - m_j)(X_i - m_i)\big) = K_{ji}$$

De même, une matrice symétrique A est dites définie positive si

$$\forall x \in \mathbb{R}^d \backslash \{0\},\ x^T A x > 0$$

Elle est dite semi-définie positive si

$$\forall x \in \mathbb{R}^d,\ x^T A x \geq 0$$

Il existe des interprétations équivalentes (qui viennent du théorème spectral).

Montrons que K est semi-définie positive. Soit $a \in \mathbb{R}^d$, on considére la forme quadratique $a^T K a$ et on rappelle que X est gaussien donc $X \in L^2$.

$$a^T K a = a^T \mathbb{E}\big((X - m)(X - m)^T\big) a = \mathbb{E}\big(a^T (X - m)(X - m)^T a\big) = \mathbb{E}\big((a^T (X - m))^2\big)$$

car $a^T (X - m)(X - m)^T a = \big((X - m)^T a\big)^T \big((X - m)^T a\big)$ est un scalaire et $X \in L^2$.
Maintenant, $(a^T (X - m))^2 \geq 0$ presque sûrement. Si bien que sont espérance est ≥ 0 et donc K est semi-définie positive.

Il y a ainsi un cas d'égalité.

$$a^T K a = 0 \iff \mathbb{E}\big((a^T (X - m))^2\big) = 0 \iff a^T (X - m) = 0 \text{ presque sûrement}$$

i.e.

$$\ker(K) = \{a \in \mathbb{R}^d | a^T K a = 0\} = \{a \in \mathbb{R}^d | a^T (X - m) = 0 \text{ p.s.}\}$$

Comme développement, on peut aussi voir ce que l'on peut faire avec raisonnement spectral (i.e. théorème spectral et valeurs propres positives), ce que l'on peut faire avec les racines carrées (Cholesky etc), lien avec loi multidimensionnelle..

Quand on parle de vecteur gaussien standard de dimension d, on notera

$$X \sim \mathcal{N}_d(m, K)$$

Exemple 8.3 (Exemple canonique du vecteur gaussien standard). On considère un vecteur aléatoire $Z = (Z_1, ..., Z_d)^T : \Omega \to \mathbb{R}^d$ dont les composantes Z_i sont indépendantes et telles que $Z_i \sim \mathcal{N}(0, 1)$. Autrement dit, chaque Z_i a pour densité

$$f_{Z_i}(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right),\ x \in \mathbb{R}$$

Puisque les Z_i sont indépendantes

$$f_Z(x_1, ..., x_d) = \prod_{i=1}^{d} f_{Z_i}(x_i) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^{d} x_i^2\right) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2} ||x||^2\right)$$

Ainsi, Z est un vecteur gaussien de dimension d. On a

$$\mathbb{E}(Z) = (\mathbb{E}(Z_1), ..., \mathbb{E}(Z_d))^T = 0$$

Et,

$$K_Z = \mathbb{E}\big((Z - \mathbb{E}(Z))(Z - \mathbb{E}(Z))^T\big) = \mathbb{E}(ZZ^T)$$

Or les composantes sont indépendantes (produit des espérance) et centrées (espérance vaut 0), donc tous les coefficients en dehors de la diagonale de la matrice de covariance sont nuls. Si bien que

$$K_Z = \text{diag}(\mathbb{E}(Z_1^2), ..., \mathbb{E}(Z_d^2))$$

Comme les composantes sont telles que $Z_i \sim \mathcal{N}(0,1)$ alors les moments d'ordres 2 sont égaux à 1, c'est-à-dire

$$K_Z = I_d$$

Finalement,

$$Z \sim \mathcal{N}_d(0, I_d)$$

i.e. Z suit la loi normale multidimensionnelle standard.

On peut aussi justifier le fait que le vecteur Z soit gaussien en disant que pour tout $a \in \mathbb{R}^d$

$$a^T Z = \sum_{i=1}^{d} a_i Z_i$$

est une combinaison linéaire de variables aléatoires normales indépendantes, qui est donc encore une gaussienne

$$a^T Z \sim \mathcal{N}(0, a^T I_d a) = \mathcal{N}(0, ||a||^2)$$

On peut aussi faire une petite digression sur les autres quantités à calculer

$$\varphi_Z(t) = \mathbb{E}\bigg(\exp\big(et^T Z\big)\bigg),\ t \in \mathbb{R}^d$$

Par indépendance

$$\varphi_Z(t) = \prod_{i=1}^{d} \mathbb{E}\big(it_i Z_i\big) = \prod_{i=1}^{d} \exp\bigg(-\frac{t_i^2}{2}\bigg) = \exp\bigg(-\frac{1}{2}||t||^2\bigg)$$

Ce qui est la forme canonique de la fonction caractéristique d'un vecteur gaussien standard. Aussi,

8.2 Processus Stochastiques

Dans ce chapitre, on présentera la notion générale d'un processus stochastique. Pour cela on décrit d'abord les lois des processus, leurs propriétés, les trajectoires des processus ainsi que la notion de convergence faible des processus.

Définition 8.2.1 (Processus stochastique)**.** Un processus stochastique $X = (X_t)$ est une famille de variables aléatoires X_t indexée par un ensemble T.

En général $T = \mathbb{R}$ ou $\mathbb{R}_+$ et on considère que le processus est indexé par le temps t. Si T est un ensemble fini, le processus est alors un vecteur aléatoire. Si $T = \mathbb{N}$ alors le processus est une suite de variables aléatoires et plus généralement si $T \subseteq \mathbb{Z}$ alors le processus est dit discret. Pour $T \subseteq \mathbb{R}^d$ on parle de champ aléatoire (d'un drap pour $d = 2$).

Il est important de comprendre les objet qu'on manipule et plus spécifiquement le fait qu'un processus dépende de 2 paramètres : $X_t(\omega)$ dépend de t et de l'aléatoire $\omega \in \Omega$.

- Pour t fixé, $\omega \in \Omega \mapsto X_t(\omega)$ est une variable aléatoire de l'espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$;
- Pour $\omega \in \Omega$ fixé, $t \mapsto X_t(\omega)$ est une fonction à valeurs réelles appelée tajectoire du processus. C'est un enjeu de savoir si un processus admet des trajectoires mesurables, continues, dérivables ou encore plus régulières.

Dans la suite, on prendra $T = \mathbb{R}_+$ pu $[0, 1]$.

8.2.1 Loi d'un processus

Comme on vient de le constater, on peut voir le processus stochastique X comme une application aléatoire

$$\begin{array}{rccc} X & : & \Omega & \to & \mathbb{R}^T \\ & & \omega & \mapsto & (X_t(\omega))_{t \in T} \end{array}$$

Autrement dit, chaque réalisation du processus est une fonction $t \mapsto X_t(\omega)$.
Cependant, l'espace T est souvent énorme voir non-mesurable de manière naturelle ($T = \mathbb{R}_+$), donc on ne peut pas toujours définir une loi de probabilité sur cet espace.

On a donc l'idée de décrire les processus par ses projections finies.

Définition 8.2.2 (Loi fini-dimensionnelles)**.** On appelle loi fini-dimensionnelles d'un processus l'ensemble des lois

$$\left\{ \mathcal{L}(X_{t_1}, ..., X_{t_p}) | t_1, ..., t_p \in T, p \in \mathbb{N}^* \right\}$$

Un processus X est à valeurs dans $\mathbb{R}^T$. On munit $\mathbb{R}^T$ de la tribu cylindrique $\sigma(Cyl)$ engendrée par la famille des cylindres

$$Cyl = \left\{ \{x : T \to \mathbb{R} | x(t_1) \in A_1, ..., x(t_p) \in A_p\} \middle| A_1, ..., A_p \in \mathcal{B}(\mathbb{R}), p \in \mathbb{N}^* \right\}$$

Il s'agit de la tribu sur $\mathbb{R}^T$ rendant mesurables les applications coordonnées.

Enfait, pour un processus (X_t), $t \in T$ et pour tout choix d'indices finis $t_1, ..., t_p \in T$, on considère le vecteur aléatoire $(X_{t_1}, ..., X_{t_p}) \in \mathbb{R}^p$. Sa loi est une probabilité sur $(\mathbb{R}^p, \mathcal{B}(R^p))$ notée

$$\mu_{t_1,...,t_p}(A) = \mathbb{P}\big((X_{t_1}, ..., X_{t_n}) \in A\big), \ A \in \mathcal{B}(\mathbb{R}^p)$$

On appelle lois fini-dimensionnelles du processus l'ensemble des mesures

$$\left\{\mu_{t_1,...,t_p} | p \in \mathbb{N}^*,\ t_1, ..., t_p \in T\right\}$$

Les lois fini-dimensionnelles sont les observations cohérentes du processus. Elles contiennent toute l'information probabiliste accessible sans passer par la complexité de $\mathbb{R}^T$. Elles décrivent toutes les dépendances statistiques entre un nombre fini d'instants. Autrement dit, on verra que connaître la loi de tout vecteur $(X_{t_1}, ..., X_{t_p})$ pour tout p et tout choix de $t_1, ..., t_p$, c'est connaître la loi du processus.

La justification que cette approche fonctionne, i.e. que les lois fini-dimensionnelles de X définissent une loi sur $(\mathbb{R}^T, \sigma(Cyl))$ est garantie par le théorème d'extension de Kolmogorov

Théorème 8.2.1 (Extension de Kolmogorov). *Soit $\mathcal{Q} = \{Q_{t_1,...,t_p}\}$ une famille de lois fini-dimensionnelles vérifiant les conditions de compatibilité*

- *(Symétrie) Si $s = (t_{i_1}, ..., t_{i_p})$ est une permutation de $t = (t_1, ..., t_p)$, alors pour tout $A_i \in \mathcal{B}(\mathbb{R})$,*

$$Q_t(A_1 \times . \times A_p) = Q_s(A_{i_1}, ..., A_{i_p})$$

- *(Compatibilité marginale) Si $t = (t_1, ..., t_p)$ et $s = (t_1, ..., t_{p-1})$ et $A \in \mathcal{B}(\mathbb{R}^{p-1})$ alors*

$$Q_t(A \times \mathbb{R}) = Q_s(A)$$

  i.e. les marginales de la loi jointe doivent coïncider avec les lois de plus petite dimension. Si on oublie certaines composantes, la loi marginale correspond bien à la loi de dimension inférieure. D'une manière formelle, pour tout $m < n$, tout $t_1, ..., t_n$ et toute projection $\pi : \mathbb{R}^n \to \mathbb{R}^m$,

$$Q_{(t_1,...,t_m)}(A) = Q_{(t_1,...,t_n)}\big(\pi^{-1}(A)\big),\ \forall A \in \mathcal{B}(\mathbb{R}^m)$$

Alors dans ces conditions, il existe une mesure de probabilité $\mathbb{P}$ sur $(\mathbb{R}^T, \sigma(Cyl))$ qui admet $\mathcal{Q}$ pour famille de lois fini-dimensionnelles.

Démonstration : Admis. □

Comme les lois fini-dimensionnelles d'un processus X satisfont immédiatement les relations de compatibilité, le théorème d'extension de Kolmogorov permet effectivement de considérer la loi $\mathbb{P}_X$ d'un processus X sur $(\mathbb{R}^T, \sigma(Cyl))$.
Vérifions que c'est le cas. Soit $t_1, ..., t_n \in T$ et notons $Y_n = (X_{t_1}, ..., X_{t_n})$. On considère la projection

$$\begin{array}{rccc} \pi_{m,n} & : & \mathbb{R}^n & \to & \mathbb{R}^m \\ & & (x_1, ..., x_n) & \mapsto & (x_1, ..., x_m) \end{array}$$

Alors, $(X_{t_1}, ..., X_{t_m}) = \pi_{m,n}(X_{t_1}, ..., X_{t_n}) = \pi_{m,n}(Y_n)$. Si bien que pour tout borélien $A \subset \mathbb{R}^m$

$$Q_{(t_1,...,t_m)}(A) = \mathbb{P}\big((X_{t_1}, ..., X_{t_m}) \in A\big) = \mathbb{P}(\pi_{m,n}(Y_n) \in A) = \mathbb{P}(Y_n \in \pi_{m,n}^{-1}(A))$$

$$= Q_{(t_1,...,t_n)}(\pi_{m,n}^{-1}(A))$$

C'est la propriété de compatibilité marginale. La propriété de symétrie est automatique.

Ces 2 conditions ne sont donc pas si mystérieuse, elles expriment simplement que les projections de la même loi sur des sous-ensembles de coordonnées donnent bien les marginales correspondantes. Quand les lois finies proviennent vraiment d'un processus défini sur un espace de probabilité, tout cela découle automatiquement des propriétés fondamentales de la probabilité. En revanche dans le cas du théorème de Kolmogorov, c'est la direction inverse car on a pas de processus préexistant, on doit donc imposer ces conditions pour pouvoir recoller les morceaux et construire une loi globale cohérente sur $\mathbb{R}^T$.

Voyons maintenant un résultat très important concernant ces lois fini-dimensionnelles.

Proposition 8.2.1. *La loi $\mathbb{P}_X$ d'un processus stochastique X est entièrement caractérisé par ses lois fini-dimensionnelles.*

Démonstration : On considère deux processus $X^{(1)}$ et $X^{(2)}$ partageant les mêmes lois fini-dimensionnelles et on montre que leur loi $\mathbb{P}_1$ et $\mathbb{P}_2$ sont égales sur $(\mathbb{R}^T, \sigma(Cyl))$. On remarque que l'ensemble Cyl est un π-système (stable par intersection finie). Notons

$$\mathcal{M} = \{A \in Cyl | \mathbb{P}_1(A) = \mathbb{P}_2(A)\}$$

Il s'agit d'une classe monotone et $Cyl \subset \mathcal{M}$ (puisque les deux processus ont les mêmes lois fini-dimensionnelles). Le théorème des classes monotones assure alors que $\sigma(Cyl) \subset \mathcal{M}$. □

Il y a plusieurs façons pour des processus stochastiques d'être égaux.

Définition 8.2.3 (Egalité des processus)**.** .

- Deux processus X et Y ont la même loi s'ils ont les mêmes lois fini-dimensionnelles pour tout p et tout $t_1, ..., t_p \in T$.
- On dira que Y est une version (ou une modification) du processus X si pour tout $t \in T$ on a $\mathbb{P}(X_t = Y_t) = 1$.
- Deux processus X et Y sont dit indistinguables s'il existe $N \in \mathcal{F}$ négligeable tels que, pour tout $\omega \notin N$, on a $X_t(\omega) = Y_t(\omega)$ pour tout $t \in T$. (Egalité par trajectoire presque sûrement)

Proposition 8.2.2.

$$\textit{Indistinguable} \Rightarrow \textit{Modification} \Rightarrow \textit{Même lois fini-dimensionnelles}$$

On peut introduire quelques exemples élémentaires pour se familiariser avec les notions et aussi pour produire des contre-exemples concernant les réciproques de ces implications.

Exemple 8.4. Soit $N \sim \mathcal{N}(0,1)$. Posons pour tout t, $X_t = N$ et $Y_t = -N$. Alors d'une part, X et Y ont les mêmes loi fini-dimensionnelles. En effet, comme la loi normale standard est symétrique, N et $-N$ ont la même loi, il en va donc de même pour X et Y et cela implique donc que les deux

processus ont les mêmes lois fini-dimensionnelles. On peut aussi le prouver plus formellement : Fixons des temps $t_1, ..., t_n$. Les vecteurs finis sont

$$(X_{t_1}, ..., X_{t_n}) = (N, ..., N) \text{ et } (Y_{t_1}, ..., Y_{t_n}) = (-N, ..., -N)$$

Ces vecteurs prennent donc toujours des valeurs sur la diagonale

$$D = \{(x, ..., x) | x \in \mathbb{R}\} \subset \mathbb{R}^n$$

Pour tout borélien $A \subset \mathbb{R}^n$ on a donc en posant $D_A = \{x \in \mathbb{R} | (x, ..., x) \in A\}$,

$$\mathbb{P}\big((X_{t_1}, ..., X_{t_n})\big) = \mathbb{P}(N \in D_A) \text{ et } \mathbb{P}\big((Y_{t_1}, ..., Y_{t_n}) \in A\big) = \mathbb{P}(-N \in D_A)$$

Or, la loi de N étant symétrique (loi normale standard),

$$\mathbb{P}(N \in D_A) = \mathbb{P}(-N \in D_A)$$

Ainsi, les lois fini-dimensionnelles coïncident.
Cependant, on a

$$\mathbb{P}(X_t = Y_t) = \mathbb{P}(N = -N) = \mathbb{P}(2N = 0) = \mathbb{P}(N = 0)$$

Or, sous la loi normale standard, N a une densité continue et pour une variable aléatoire à densité, $\mathbb{P}(N = a) = 0$ pour tout a. Si bien que

$$\mathbb{P}(X_t = Y_t) = 0$$

Ce qui prouvent que même lois fini-dimensionnelles n'impliquent pas que les deux processus sont une version l'un de l'autre.

Exemple 8.5. On se place sur l'espace de probabilité $([0,1], \mathcal{B}([0,1]), \lambda)$ et on pose $T = [0,1]$. On considère la diagonale D de $[0,1]$, ie.

$$D = \{(t, \omega) \in [0, 1^2 | t = \omega]\}$$

et on définit

$$X_t(\omega) = 0, \forall (t, \omega), \quad Y_t(\omega) = \mathbb{1}_D(t, \omega) = \mathbb{1}_{t=\omega}(t, \omega)$$

Maintenant, $\lambda(\{t = \omega\}) = 0$ et comme $X_t(\omega) = 0$ alors $\mathbb{P}(X_t = Y_t) = 1$. Si bien que X et Y sont version l'un de l'autre.
Cependant, les deux processus ne sont pas indistinguables. En effet, être indistinguable signifie qu'il existe un ensemble de probabilité 1 sur lequel les trajectoires coïncident pour tous les t. Ici pour un ω fixé, $Y_t(\omega) = 1 \Leftrightarrow t = \omega$. Si bien que la trajectoire de Y associé à cet ω vaut 1 et vaut 0 pour tous les autres t. En particulier, pour tout ω on a $Y_\omega(\omega) = 1$ donc pour aucun ω, la trajectoire $t \mapsto Y_t(\omega)$ n'est égale à la trajectoire identiquement nulle. Donc les deux processus ne sont pas indistinguables.
Enfait ici, comme D est de dimension 1 dans un carré de dimension 2, sa mesure de Lebesgue est nulle ce qui implique que les processus sont égaux presque partout sur ce carré. Cependant l'indistinguabilité exige une condition plus forte (égalité pour tous t simultanément sur un ensemble de probabilité 1).

On peut aller un peu plus loin das l'explication en disant que la non-indistinguabilité de X et Y vient du fait que les trajactoires de X sont continues tandis que celles de Y ne le sont pas. C'est ce qu'il manque pour avoir une réciproque :

Proposition 8.2.3. *Soient T séparable (i.e. T contient une partie dense dénombrable) et X, Y des versions avec des trajectoires continues presque sûrement, alors ils sont indistinguables.*

Démonstration : On choisit D une partie dense dénombrable dans T. Alors, pour tout $t \in D$ on a $\mathbb{P}(X_t = Y_t) = 1$ car $D \subset T$. De plus comme D est dénombrable, l'ensemble $A = \{X_t = Y_t | t \in D\} \in \mathcal{F}$ est de probabilité 1. L'ensemble $B = \{X$ et Y sont à trajectoires continues$\}$ est aussi de probabilité 1 par hypothèse. Soit $\omega \in A \cap B$ tel que les trajectoires soient continues. On a $\mathbb{P}(A \cap B) = 1$ et pour $w \in A \cap B$,

- Si $t \in D$, on a $X_t = Y_t$;
- Sinon, il existe $t_n \in D$ avec $t_n \to t$ par densité. On a donc $X_{t_n} = Y_{t_n}$ et par continuité des trajectoires pour $\omega \in B$, $X_{t_n} \to X_t$ et $Y_{t_n} \to Y_t$. On a donc $X_t = Y_t$ sur $A \cap B$ par unicité de la limite.

Finalement, pour $\omega \in A \cap B$, $X_t = Y_t$ pour tout $t \in T$, ce qui signifie que les deux processus sont indistinguables. □

Exemples de propriétés en loi des processus :

Il existe de nombreuses classes de processus particuliers : les processus de Markov (voir chaînes de Markov et Modèles Aléatoires), les Martingales, les processus gaussiens, les processus de Poisson (voir Modèles Aléatoires), les processus stables ou encore les processus de Lévy (qui contiennent les 3 exemples précédents). Ces types de processus sont caractérisés par des propriétés remarquables de leurs lois fini-dimensionnelles. On donne ici quelques exemples de telles propriétés en loi des processus.

Définition 8.2.4. .

- Un processus est dit (strict) stationnaire si pour tout $h \geq 0$, $(X_{t+h}) =^{\mathcal{L}} (X_t)$ ne dépend pas de $h > 0$. C'est-à-dire que pour tout $h > 0$ et tout $t_1, ..., t_p \geq 0$, on a

$$(X_{t_1+h}, ..., X_{t_p+h}) =^{\mathcal{L}} (X_{t_1}, ..., X_{t_p})$$

- Un processus est dit à accroissements stationnaires si la loi des accroissements $X_{t+h} - X_t$ ne dépend pas de $t > 0$, i.e.

$$X_{t+h} - X_t =^{\mathcal{L}} X_h$$

- Un processus est dit à accroissements indépendants si pour tout $p \geq 1$ et pour tout $0 < t_1 < ... < t_p$, les variables aléatoires X_{t_1}, $X_{t_2} - X_{t_1}$, ..., $X_{t_p} - X_{t_{p-1}}$ sont indépendants.

Un des exemples les plus utiles, c'est la marche aléatoire.

Exemple 8.6. Ici, $T = \mathbb{N}$. Soit (X_n) une suite de variables aléatoires indépendantes. On considère $S_n = X_1 + ... + X_n$ le processus discret des sommes partielles. On parle de marche aléatoire. Alors (S_n) est un processus à accroissements indépendants. Montrons cela.
On veut montrer que les accroissements

$$\Delta_j = S_{n_j} - S_{n_{j-1}} = \sum_{i=n_{j-1}+1}^{n_j} X_i$$

pour une suite d'indices $0 = n_0 < n_1 < ... < n_m$ sont indépendants entre eux. Pour cela, posons d'abord $\mathcal{G}_i = \sigma(X_i)$. L'indépendance des X_i par hypothèse équivaut à l'indépendance des $\big(\mathcal{G}_i\big)_{i\geq 1}$. C'est-à-dire, pour tout choix d'entiers distincts $i_1, ..., i_r$ et d'ensembles mesurables $A_k \in \mathcal{G}_{i_k}$

$$\mathbb{P}\Big(\bigcap_{k=1}^{r} A_k\Big) = \prod_{k=1}^{r} \mathbb{P}(A_k)$$

Pour un bloc d'indices $B_j = \{n_{j-1} + 1, ..., n_j\}$, on définit $\mathcal{H}_i = \sigma(X_i | i \in B_j)$. Comme les $\mathcal{G}_i$ sont indépendantes, alors les $\mathcal{H}_i$ le sont aussi, chacune étant engendrée par des $\mathcal{G}_i$ d'indices disjoints.
Maintenant, chaque accroissements Δ_j est $\mathcal{H}_j$-mesurable. D'où l'indépendance des accroissements.
Maintenant, si les variables aléatoires (X_i) sont aussi de même loi (i.e. les X_i sont iid), alors les accroissements sont indépendants et stationnaires. Il y a plusieurs moyens de le montrer. La plus rapide consiste à utiliser les fonctions caractéristiques,

$$\forall t \in \mathbb{R}^d,\ \varphi_{X_i}(t) = \varphi_X(t) = \mathbb{E}\big(e^{i\langle t, X_i\rangle}\big)$$

Par indépendance des X_i, on a

$$\varphi_{S_n}(t) = \big(\varphi_X(t)\big)^n$$

et, $S_{n+k} - S_k = X_{n+k} + ... + X_{k+1}$ est une somme à n termes des variables aléatoires iid, d'où

$$\varphi_{S_{n+k}-S_k}(t) = \big(\varphi_X(t)\big)^n$$

Si bien que, $\varphi_{S_{n+k}-S_k} = \varphi_{S_n}$, d'où l'égalité en loi. Cependant, on ne comprend pas très bien ce qu'il se passe.
Une autre manière de faire consisterait à dire que comme les (X_i) sont i.i.d., alors le vecteurs aléatoire $(X_{k+1}, ..., X_{k+l})$ a la même loi que le vecteur $(X_1, ..., X_l)$. Or la somme est une application mesurable, si bien que

$$\sum i = k + 1^{k+l} X_i =^{\mathcal{L}} \sum_{i=1}^{l} X_i$$

D'où le résultat.

8.2.2 Régularité des trajectoires

D'après les exemples ci-dessus, les versions d'un processus stochastique n'ont pas toujours la même régularité de leurs trajectoires. Aussi, il est intéressant de chercher si un processus X admet des versions $\tilde{X}$ dont les trajectoires ont de "bonnes" propriétés de régularité et d'avoir des conditions le garantissant.

Dans cette section, on s'attache à trouver des versions à trajectoires continues d'un processus.

Théorème 8.2.2 (Kolmogorov-Centsov). *Soit $(X_t)_{t\in T}$ un processus indexé par un intervalle $T \subset \mathbb{R}$ de $\mathbb{R}$ et à valeurs dans un espace métrique complet (E,d).*
On suppose qu'il existe $a,b,C>0$ vérifiant

$$\forall s,t \in T,\ \mathbb{E}\big(d(X_t,X_s)^a\big) \leq C|t-s|^{1+b}$$

Alors, il existe une version $\tilde{X}$ de X dont les trajectoires sont localement höldériennes d'exposant γ pour tout $\gamma \in]0,\frac{b}{a}[$, i.e.

$$\forall s,t \in T,\ d\big(\tilde{X}_t(\omega),\tilde{X}_s(\omega)\big) \leq C_\gamma(\omega)|t-s|^\gamma$$

En particulier, $\tilde{X}$ est une version continue de X.

L'hypothèse de ce théorème donne un contrôle momentané de l'ordre a sur les incréments du processus. Si ces incréments sont "assez petits en moyenne" (i.e. que le moment d'ordre a décroît plus vite que $|t-s|$), on peut construire une version du processus dont les trajectoires sont presque sûrement régulières.
Le théorème transforme donc une régularité en loi (via les moments) en une régularité presque sûre (des trajectoires).

Avant de montrer ce théorème, on aura besoin d'un lemme central. Ce lemme fournit le pont entre le contrôle probabiliste des incréments (via les moments) et la construction d'une trajectoire Höldérienne presque sûre.

Lemme 8.2.1. *Soit $(X_t)_{t\in[0,1]}$ un processus stochastique à valeurs dans un espace métrique complet (E,d). On suppose qu'il existe $a,b,C>0$ tels que*

$$\forall s,t \in [0,1],\ \mathbb{E}\big(d(X_t,X_s)\big) \leq C|t-s|^{1+b}$$

Soit $D_n = \left\{ \frac{k}{2^n} | k \in [0,2^n] \right\}$ l'ensemble des points dyadiques de pas 2^{-n}.
Alors, il existe un ensemble négligeable $N \subseteq \Omega$ tel que, pour tout $\omega \notin N$, il existe une constante $C(\omega) < \infty$ vérifiant

$$\forall s,t \in D = \bigcup_n D_n,\ \forall \gamma \in \left]0,\frac{b}{a}\right[,\ d\big(X_t(\omega),X_s(\omega)\big) \leq C(\omega)|t-s|^\gamma$$

Autrement dit, presque sûrement les trajectoires du processus sont Höldériennes d'ordre arbitrairement proche de $\frac{b}{a}$ sur les points dyadiques.

<u>*Démonstration :*</u> D'après l'inégalité de Markov sur l'hypothèse du lemme (et du théorème), pour $a>0$ et $s,t\in[0,1]$, $u>0$

$$\mathbb{P}\big(d(X_s,X_t)\geq u\big) \leq \frac{\mathbb{E}\big(d(X_s,X_t)^a\big)}{u^a} \leq \frac{C|t-s|^{1+b}}{u^a}$$

En appliquant cette inégalité avec $s=\frac{i-1}{2^n}$, $t=\frac{i}{2^n}$ (avec $i\in[1,2^n]$) et $u=2^{-n\gamma}$, on a

$$\mathbb{P}\big(d(X_{(i-1)2^{-n}},X_{i2^{-n}})\geq 2^{-n\gamma}\big) \leq C2^{na\gamma}2^{-(1+b)n}$$

En sommant sur $i \in [1, 2^n]$:

$$\sum_{i=1}^{2^n} \mathbb{P}\big(d(X_{(i-1)2^{-n}}, X_{i2^{-n}}) \geq 2^{-n\gamma}\big) = \mathbb{P}\bigg(\bigcup_{i=1}^{2^n} \{d(X_{(i-1)2^{-n}}, X_{i2^{-n}}) \geq 2^{-n\gamma}\}\bigg)$$

$$\leq \sum_{i=1}^{2^n} C 2^{na\gamma} 2^{-(1+b)n} = 2^n C 2^{na\gamma} 2^{-(1+b)n} = C 2^{-n(b-a\gamma)}$$

Maintenant, comme $b - a\gamma > 0$, en sommant sur $n > 0$,

$$\sum_{n=0}^{\infty} \mathbb{P}\bigg(\bigcup_{i=1}^{2^n} \{d(X_{(i-1)2^{-n}}, X_{i2^{-n}})\}\bigg) < +\infty$$

Si bien que, d'après le lemme de Borel-Cantelli, il existe presque sûrement $n_0(\omega) \in \mathbb{N}$ tel que dès que $n \geq n_0(\omega)$ pour tout $i \in [1, 2^n]$, on a

$$d\big(X_{(i-1)2^{-n}}, X_{i2^{-n}}\big) \leq 2^{-n\gamma}$$

Et donc a fortiori que

$$K_\gamma(\omega) = \sup_{n \geq 1} \bigg(\sup_{1 \leq i \leq 2^n} \frac{d\big(X_{(i-1)2^{-n}}, X_{i2^{-n}}\big)}{2^{-n\gamma}} \bigg) < \infty$$

On obtient alors le résultat du Lemme avec

$$C_\gamma(\omega) = 2^{\gamma+1} \frac{1 - 2^{-\gamma} + 2^{1-\gamma}}{1 - 2^{-\gamma}} K_\gamma(\omega)$$

En effet, considérons $s, t \in D$ avec $s < t$. Soit $p \geq 1$ tel que $2^{-(p+1)} < t - s < 2^{-p}$. Il existe $m \geq 1$ tel qu'on puisse écrire $s, t \in D$ sous la forme, avec $\varepsilon_j, \varepsilon_j^{'} \in \{0, 1\}$:

$$s = k2^{-p} + \varepsilon_0 2^{-p} + \varepsilon_1 2^{-(p+1)} + ... + \varepsilon_m 2^{-(p+m)}$$

$$t = k2^{-p} + \varepsilon_0^{'} 2^{-p} + \varepsilon_1^{'} 2^{-(p+1)} + ... + \varepsilon_m^{'} 2^{-(p+m)}$$

On note pour $j \in [0, m]$:

$$s_j = k2^{-p} + \varepsilon_0 2^{-p} + \varepsilon_1 2^{-(p+1)} + ... + \varepsilon_m 2^{-(p+j)}$$

$$t_j = k2^{-p} + \varepsilon_0^{'} 2^{-p} + \varepsilon_1^{'} 2^{-(p+1)} + ... + \varepsilon_m^{'} 2^{-(p+j)}$$

De sorte que $s = s_m$ et $t = t_m$. Par inégalité triangulaire on a alors

$$d\big(X_s, X_t\big) = d\big(X_{s_m}, X_{t_m}\big)$$

$$\leq d(X_{s_0}, X_{t_0}) + \sum_{j=1}^{m} d(X_{s_{j-1}}, X_{s_j}) + \sum_{j=1}^{m} d(X_{t_{j-1}}, X_{t_j})$$

Puis, comme $\sum_{j=1}^{m} 2^{-(p+j)\gamma} \leq \frac{2^{-p\gamma}2^{-\gamma}}{1-2^{-\gamma}}$, il vient

$$\leq K_\gamma(\omega)2^{-p\gamma} + K_\gamma(\omega)2^{-p\gamma}\frac{2^{-\gamma}}{1-2^{-\gamma}} + K_\gamma(\omega)2^{-p\gamma}\frac{2^{-\gamma}}{1-2^{-\gamma}}$$

et comme $2^{-p} \leq 2(t-s)$, il vient

$$\geq C_\gamma(\omega)(t-s)^\gamma$$

□

Là, on pourrait se dire que les nombres dyadiques sortent un peu de nul part et que leur utilisation n'est pas vraiment intuitive : c'est partiellement le cas. Les nombres dyadiques ont un rôle central dans beaucoup de preuves en analyse et en probabilités (notamment dans la théorie de la régularité des trajectoires, rtingales, et les approximations de fonctions). Mais pourquoi ?

C'est un ensemble dénombrable et dense dans $[0,1]$, ce qui est idéal pour obtenir des résultat presque sûrs : on ne peut pas vérifier une propriété sur une infinité non-dénombrable de points, mais on le peut sur une suite dénombrable dense puis prolonger par continuité.

Démonstration : (Kolmogorov-Centsov)
Nous supposons que $T = [0,1]$. Si T est non-borné (par exemple $\mathbb{R}_+$), on peut appliquer le cas borné à $T = [0,1], [1,2], [2,3]...$ et on trouver encore que X a une modification continue définie sur T, qui est localement höldérienne d'exposant γ pour tout $\gamma \in \left]0, \frac{b}{a}\right[$. Pour simplifier, on prend $T = [0,1]$ dans la suite.
Il suffit de montrer que pour $\gamma \in \left]0, \frac{b}{a}\right[$ fixé, X a une version dont les trajectoires sont höldériennes d'exposant γ. En effet, on appliquera alors ce résultat à une suite $\gamma_n \to \frac{b}{a}$ en observant que les processus obtenus sont des versions continues du même processus X donc indistinguables.
On note D l'ensemble (dénombrable) des nombres dyadiques $t \in [0,1[$ qui s'écrivent sous la forme

$$t = \sum_{k=1}^{p} \varepsilon_k 2^{-k},\ \varepsilon \in \{0,1\},\ 1 \leq k \leq p$$

D'après le lemme que l'on a montré précédemment, la fonction $t \mapsto X_t(\omega)$ est presque sûrement γ-höldérienne sur D, donc uniformément continue sur D. Comme (E,d) est complet, il existe presque sûrement un unique prolongement continu de cette fonction à $T = [0,1]$. Le prolongemnt reste γ-höldérien. Plus précisément, on pose pour tout $t \in T$

$$\tilde{X}_t(\omega) = \lim_{s \to t,\ s \in D} X_s(\omega)$$

sur l'ensemble presque sûr $\{\omega \in \Omega | K_\gamma(\omega) < \infty\}$ où $s \mapsto X_s(\omega)$ est γ-höldérienne sur D et on pose $\tilde{X}_t(\omega) = x_0$ sur l'ensemble $\{\omega \in \Omega | K(\omega) = \infty\}$ négligeable x_0 est un point ficé quelconque de E. Par construction, le prcessus $\tilde{X}$ a alors des trajectoires höldériennes d'exposant γ sur T.

Il reste à voir que $\tilde{X}$ est bien une version de X. Or l'hypothèse avec l'inégalité de Markov

$$\mathbb{P}\big(d(X_s, X_t) \geq \varepsilon\big) \leq \frac{\mathbb{E}\big(d(X_s, X_t)^a\big)}{\varepsilon^a} \leq \frac{|t-s|^{1+b}}{\varepsilon^a}$$

Ainsi pour tout $t \in T$ fixé

$$X_s \to^{\mathbb{P}} X_t$$

Comme par construction $X_s \to \tilde{X}_t$, on conclut que $X_t = \tilde{X}_t$ presque sûrement par unicité presque sûrement de la limite en probabilité.□

On a utilisé la notion de fonctions höldérienne que l'on a appliqué aux processus stochastiques. On va se faire un rapide topo sur les fonctions höldériennes.

Définition 8.2.5 (Fonctions α-höldérinnes)**.** Soit $I \subseteq \mathbb{R}^d$ (souvent $I = [0,1]$) et $f : I \to \mathbb{R}^m$. On dit que f est höldérienne d'ordre $\alpha > 0$ sur I s'il existe une constante $C > 0$ telle que

$$\forall x, y \in I,\ ||f(x) - f(y)|| \leq C||x-y||^\alpha$$

Le plus petit C pour lequel cette inégalité est vérifiée s'appelle la constante de Hölder de f, notée souvent $[f]_{\mathcal{C}^{0,\alpha}}$.

On remarque que

- Si $\alpha = 1$ alors f est Lipschitzienne.
- Si $0 < \alpha < 1$, on parle de continuité Höldérienne stricte (plus α est grand plus f est régulière) et f est alors uniformément continue (car $||x-y||^\alpha \to 0$ quand $x \to y$).
- Si $\alpha > 1$ alors f est nécessairement constante (Preuve classique ; Passer par la différentielle pour un espace vectoriel normé et par un partitionnement pour un espace métrique. On peut aussi ramener le cas métrique au cas vectoriel en embarquant isométriquement l'espace métrique dans un Banach, via l'injection de Kuratowski-Fréchet dans $l^\infty(E)$, et appliquer la preuve différentielle au composé - la dérivée peut donc être simulée même sans être initialement dans un espace vectoriel).

Exemple 8.7. Soit $f : x \mapsto \sqrt{x}$ sur $I = [0,1]$. Soient $x, y \geq 0$, on peut supposer sans perte de généralités que $x \geq y$. Alors

$$x - y = (\sqrt{x} - \sqrt{y})(\sqrt{x} + \sqrt{y})$$

comme $\sqrt{x} + \sqrt{y} \geq \sqrt{x} - \sqrt{y} \geq 0$,

$$x - y \geq (\sqrt{x} - \sqrt{y})^2$$

Donc,

$$\sqrt{|x-y|} \geq |\sqrt{x} - \sqrt{y}|$$

Autrement dit, $\sqrt{\cdot}$ est Höldérienne d'ordre ${}_alpha = \frac{1}{2}$ avec $C = 1$

$$\forall x, y \geq 0,\ |\sqrt{x} - \sqrt{y}| \leq |x-y|^{\frac{1}{2}}$$

Concernant l'optimalité, si on cherchait un ordre $\alpha > \frac{1}{2}$, prendre $y = 0$ et $x \to 0^+$ donne

$$\frac{|\sqrt{x} - 0|}{|x-0|^\alpha} = x^{\frac{1}{2}-\alpha} \to \infty$$

donc aucun $\alpha > \frac{1}{2}$ ne convient globalement : $\frac{1}{2}$ est l'exposant maximal sur tout voisinage de 0.

Intuitivement, $\sqrt{\cdot}$ est moins régulière proche de 0, ce qui empêche la Lipschitzianité mais permet la Höldérianité.

Exemple 8.8. Soit $g : x \mapsto x$ sur $I = \mathbb{R}$ (ou tout intervalle). Pour tout $x, y \geq 0$ on a $|x-y| = 1.|x-y|$ donc g est Lipschitzienne ce qui revient à dire qu'elle est Höldérienne d'ordre $\alpha = 1$, qui est forcément optimal. En effet, si $\alpha > 1$ alors par argument de dérivée, g serait constante, ce qui n'est évidemment pas le cas.
L'exposant 1 est le meilleur possible pour des fonctions non-contantes.

Exemple 8.9. Posons la fonction suivante

$$f(x) == \begin{cases} x \sin\left(\frac{1}{x}\right), & \text{si } x \neq 0, \\ 0, & \text{si } x = 0. \end{cases}$$

On a que

- f est continue en 0 (donc continue sur $[0,1]$) ;
- f est Höldérienne d'ordre α pour tout $0 < \alpha < 1$ (donc en particulier uniformément continue) ;
- f n'est pas Lipschitzienne (donc pas Höldérienne d'ordre 1).

En résumé,

$$\text{Lip.} \Rightarrow \text{Höldérienne } (0 < \alpha \leq 1) \Rightarrow \text{Uniformément } \mathcal{C}^0 \Rightarrow \mathcal{C}^0 \Rightarrow \mathcal{C}^0 \text{ en un point}$$

8.2.3 Convergence faible des lois de processus

On a vu qu'un processus $X = (X_t)_{t \in T}$ définit en une variable aléatoire sur $\left(\mathbb{R}^T, \sigma(Cyl)\right)$. Cependant avec de bonnes propriétés trajectorielles, on peut améliorer l'espace d'arrivée $\left(\mathbb{R}^T, \sigma(Cyl)\right)$, typiquement si X est à trajectoires continues alors X est à valeurs dans $\mathcal{C}^0(T, \mathbb{R})$ (ou si ce n'est X, une de ses versions en utilisant le théorème de Kolmogorov-Centsov). Quand $T = [0,1]$ (ou T borné), $\mathcal{C}^0(T, \mathbb{R})$ est normé par

$$||x - y||_\infty = \sup_{t \in T} |x(t) - y(t)|, \ x, y \in \mathcal{C}^0(T, \mathbb{R})$$

qui définit la topologie de la convergence uniforme. Il s'agit alors d'un espace de Banach.
Quand $T = \mathbb{R}_+$ (ou T borné), $\mathcal{C}^0(\mathbb{R}_+, \mathbb{R})$ admet pour distance

$$d(x, y) = \sum_{n=1}^{\infty} \frac{\min\{||x - y||_{\infty,n}, 1\}}{2^n}$$

où on note $||x - y||_{\infty,n} = \sup_{t \in [0,n]} |x(t) - y(t)|$ pour $x, \in \mathcal{C}^0(\mathbb{R}_+, \mathbb{R})$. Cette distance métrise la convergence uniforme sur tous les compacts.
L'intérêt de considérer X à valeurs dans $\mathcal{C}^0(T, \mathbb{R})$ (ou un autre espace fonctionnel, souvent $\mathbb{D}(\mathbb{R}_+, \mathbb{R})$ l'espace des fonctions càdlàg dit espace de Skorohod) est alors que X est à valeurs dans un espace métrique (et même souvent un espace métrique complet, séparable, dit espace polonais) et que dans ce cadre la notion de convergence faible est bien développée.

On commence par s'assurer que le processus $X = (X_t)_{t\in T}$ reste bien une variable aléatoire sur $\mathcal{C}^0(T,\mathbb{R})$

$$\begin{array}{rccc} X & : & (\Omega,\mathcal{F}) & \to & (\mathcal{C}^0(T,\mathbb{R}),\mathcal{B}(\mathcal{C}^0(T,\mathbb{R}))) \\ & & \omega & \mapsto & X(\omega) = \big(X_t(\omega)\big)_{t\in T} \end{array}$$

L'espace $\mathcal{C}^0(T,\mathbb{R})$ est muni de deux tribus naturelles : la trace de la tribu cylindrique $\sigma(Cyl)\cap \mathcal{C}^0(T,\mathbb{R})$ et sa propre tribu borélienne $\mathcal{B}(\mathcal{C}^0(T,\mathbb{R}))$. Les deux coïncident

Proposition 8.2.4. *Sur $\mathcal{C}^0(\mathbb{R}_+,\mathbb{R})$, la tribu cylindre trace coïncide avec la tribu borélienne*

$$\sigma(Cyl)\cap\mathcal{C}^0(T,\mathbb{R}) = \mathcal{B}(\mathcal{C}^0(T,\mathbb{R}))$$

Démonstration : On va raisonner par double inclusions.
D'abord on pose Π_{t_0} l'évaluation en t_0 qui prend une trajectoire continue x et renvoie $x(t_0)$. C'est la projection canonique sur la coordonnée t_0 (on note aussi souvent X_{t_0} dans la théorie des processus). D'abord on remarque que ¶$_{t_0}$ est continue sur $(\mathcal{C}^0(T\mathbb{R}_+,\mathbb{R}),d)$ puisque

$$\forall n\geq t_0,\ |\Pi_{t_0}(x)-\Pi_{t_0}(y)| = |x(t_0)-y(t_0)| \leq ||x-y||_{\infty,n}$$

Ainsi, Π_{t_0} est continue et donc mesurable sur $(\mathcal{C}^0(\mathbb{R}_+,\mathbb{R}),\mathcal{B}(\mathcal{C}^0(\mathbb{R}_+,\mathbb{R})))$. Comme tout cylindre s'écrit , pour $A_i\in\mathcal{B}(\mathbb{R})$

$$C = \bigcap_{i=1}^{p}\Pi_{t_i}^{-1}(A_i)$$

alors on a bien que $C\in\mathcal{B}(\mathcal{C}^0(\mathbb{R}_+,\mathbb{R}))$. Si bien que $Cyl\subseteq\mathcal{B}(\mathcal{C}^0(\mathbb{R}_+,\mathbb{R}))$ puis

$$\sigma(Cyl)\subseteq\mathcal{B}(\mathcal{C}^0(\mathbb{R}_+,\mathbb{R}))$$

Réciproquement si $A = \{y\in\mathcal{C}^0(\mathbb{R}_+,\mathbb{R})|\ ||x-y||_{[0,n]}<\eta\}$ est ouvert pour la converence uniforme sur les compacts, on a

$$A = \bigcap_{t\in[0,n]\cap\mathbb{Q}}\Pi_t^{-1}(]x(t)-\eta,x(t)+\eta[)$$

car la condition $|y(t)-x(t)|<\eta$ pour tout $t_in[0,n]\cap\mathbb{Q}$ est complétée pour tout $t\in[0,n]$ par continuité. L'ouvert A s'écrit alors comme une intersection (dénombrable) de cylindres, i.e. $A\in\sigma(Cyl)$. Puisque de tels A engendrent $\mathcal{B}(\mathcal{C}^0(\mathbb{R}_+,\mathbb{R}))$, on a

$$\mathcal{B}(\mathcal{C}^0(\mathbb{R}_+,\mathbb{R}))\subseteq\sigma(Cyl)$$

Finalement, d'après ce résultat, le processus X définit bien une variable aléatoire sur $\mathcal{C}^0(\mathbb{R}_+,\mathbb{R})$ et $\mathbb{P}_X$ est alors une loi sur $\mathcal{C}^0(\mathbb{R}_+,\mathbb{R})$.

Rappels sur la onvergence faible (en loi) :

On rappelle la notion de convergence faible de variables aléatoires à valeurs dans un espace métrique S. Quand S est un espace fonctionnel, typiquement $\mathcal{C}^0(T,\mathbb{R})$, la variable aléatoire est en fait un processus stochastique à trajectoires continues.

Définition 8.2.6 (Convergence faible)**.** Soit $(X_n)_{n\geq 1}$ et X à valeurs dans S espace métrique de lois $\mathbb{P}_n$ et $\mathbb{P}$, on a $X_n \Rightarrow X$ (ou $\mathbb{P}_n \Rightarrow \mathbb{P}$) si e seulement si pour toute fonction $f : S \to \mathbb{R}$ continue et bornée

$$\lim_{n\to\infty} \mathbb{E}\big(f(X_n)\big) = \mathbb{E}\big(f(X)\big)$$

Nous utiliserons les formulations équivalentes suivantes de la convergence faible.

Théorème 8.2.3 (Porte-manteau)**.** *Les assertions suivantes sont équivalentes quand $n \to +\infty$ pour des variables aléatoires X_n à valeurs dans un espace métrique S.*

1. *$X_n \Rightarrow X$, i.e. pour toute fonction $f : S \to \mathbb{R}$ continue et bornée*
$$\lim_{n\to\infty} \mathbb{E}\big(f(X_n)\big) = \mathbb{E}\big(f(X)\big)$$
2. *Pour toute fonction $f : S \to \mathbb{R}$ uniformément continue et bornée*
$$\lim_{n\to\infty} \mathbb{E}\big(f(X_n)\big) = \mathbb{E}\big(f(X)\big)$$
3. *Pour tout fermé F*
$$\limsup_{n\to\infty} \mathbb{P}(X_n \in F) \leq \mathbb{P}(X \in F)$$
4. *Pour tout ouvert G*
$$\liminf_{n\to\infty} \mathbb{P}(X_n \in G) \geq \mathbb{P}(X \in G)$$
5. *Pour tout $A \in \mathcal{B}(S)$ tel que $\mathbb{P}(X \in \overline{X}\backslash A^o) = 0$,*
$$\lim_{n\to\infty} \mathbb{P}(X_n \in A) = \mathbb{P}(X \in A)$$

Démonstration : Voir cours de Fondements des probabilités, section 1.7.4. □

Proposition 8.2.5. *Soit X et X_n des variables aléatoires à valeurs dans un espace métrique S.*

1. *La convergence en probabilité $X_n \to^{\mathbb{P}} X$ implique la convergence faible (en loi) $X_n \Rightarrow X$.*
2. ***(Continuous mapping theorem)** Soit $f : S \to S'$ une application entre espaces métriques. Alors si $X_n \Rightarrow X$ et f est continue $\mathbb{P}_X$-presque sûrement, on a aussi $f(X_n) \Rightarrow f(X)$. C'est-à-dire que la convergence faible (en loi) se conserve par les applications continues.*
3. *Si X_n, X sont des processus à trajectoires continues (i.e. $S = \mathcal{C}^0(\mathbb{R}_+, \mathbb{R})$) avec $X_n \Rightarrow X$, on a la convergence faible des lois fini-dimensionnelles :*
$$\forall p \geq 1,\ t_1, ..., t_p,\ \big(X_n(t_1), ..., X_n(t_p)\big) \Rightarrow \big(X(t_1), ..., X(t_p)\big)$$

Démonstration : On a déjà montré les 2 premiers points dans le cours Fondements des probabilité. Le troisième point est une conséquence du Continuous mapping theorem avec l'application continue

$$\begin{array}{rccc} \Pi_{t_1,...,t_p} & : & \mathcal{C}^0(T,\mathbb{R}) & \to & \mathbb{R}^p \\ & & x & \mapsto & \Pi_{t_1,...,t_p}(x) = \big(x(t_1), ..., x(t_p)\big) \end{array}$$

□

Le troisième point dit que la convergence faible de processus à trajectoires continues entraîne la convergence des lois fini-dimensionnelles. La réciproque est fausse : la convergence des lois fini-dimensionnelles n'entraîne pas la convergence faible (en loi) es processus. Il peut y avoir un phénomène de perte de masse vers l'infini comme illustré dans l'exemple suivant pour des variables aléatoires réelles.
De plus, l'affirmation est fausse pour des processus qui ne sont pas à valeurs dans $\mathcal{C}^0(T,\mathbb{R})$: par exemple pour des processus càdlàg, $X_n \Rightarrow X$ entraîne $X_n(t) \Rightarrow X(t)$ seulement si X est continu en t.

Exemple 8.10. Sur $\mathcal{C}^0([0,1],\mathbb{R})$, on considère les fonctions

$$x_n(t) = \begin{cases} 2nt & \text{si } t \in [0, \frac{1}{2n}], \\ 2 - 2nt & \text{si } t \in [\frac{1}{2n}, \frac{1}{n}]. \end{cases}$$

La suite de fonctions x_n converge simplement vers $x = 0$ (En effet, si $t = 0$ zlors pour tout n on a $x_n(0) = 0$; si $t > 0$ fixé, choisissons $N \in \mathbb{N}$ tel que $N > \frac{1}{t}$, pour tout $n \geq N$ on a $\frac{1}{n} < t$donc $t \notin [0, \frac{1}{n}]$. Par la définition ci-dessus $x_n(t) = 0$ pour tout $n \geq N$, ainsi $x_n(t) \to 0$). La convergence n'est cependant pas uniforme sur $[0,1]$ car pour chaque n la valeur maximale de x_n est atteinte en $t = \frac{1}{2n}$ et vaut

$$x_n\left(\frac{1}{2n}\right) = 2n.\frac{1}{2n} = 1$$

Si bien que $||x_n||_\infty = 1$ pour tout n, qui ne tend pas vers 0.
On considère la suite de loi $\mathbb{P}_n = \delta_{x_n}$ et $\mathbb{P} = \delta_x$ et on observe que

- On a pas $\mathbb{P}_n \Rightarrow P$ car pour la fonction continue bornée $f(y) = \sup_{t\in[0,1]} y(t)$, on a $f(x_n) = 1$, $f(x) = 0$, soit

$$\int f d\mathbb{P}_n = f(x_n) = 0 \text{ ne tend pas vers } 0 = f(x) = \int f d\mathbb{P}$$

- Pour tout $0 < t_1 < ... < t_p$ avec $\Pi_{t_1,...,t_p}(y) = \big(y(t_1), ..., y(t_p)\big)$ on a $\Pi_{t_1,...,t_p}(x) = (0, ..., 0)$ et pour $n > \frac{1}{t_1}$ aussi $\Pi_{t_1,...,t_p}(x_n) = (0, ..., 0)$, c'est-à-dire la convergence des lois fini-dimensionnelles

$$\mathbb{P}_n \Pi^{-1}_{t_1,...,t_p} \Rightarrow \mathbb{P}\Pi^{-1}_{t_1,...,t_p}$$

La convergence des lois fini-dimensionnelles n'entraîne donc pas la convergence en loi de tout le processus. Pour éviter une telle pathologie, il faut une condition supplémentaire et c'est l'objet de ce qui suit.

Equitension :

L'équitension est la condition qui empêche la perte de masse probabiliste. C'est la condition qui avec la convergence des lois fini-dimensionnelles donnera la convergence faible. (Plus de détail sur cette partie globale dans *Convergence of Probability measures* de Patrick Billingsley)

Définition 8.2.7 (Equitension)**.** Soit (P_n) une suite de mesures de probabilité sur un espace métrique. La suite est dite équitendue si pour tout $\varepsilon > 0$, il existe un compact K_ε tel que pour tout $n \geq 1$ on a $P_n(K_\varepsilon) > 1 - \varepsilon$.

En fait, l'équitension s'exprime aussi par une propriété de relative compacité (dans les bons espaces métriques : ceux qui sont séparables et complets, i.e. polonais).

Définition 8.2.8 (Relative compacité)**.** Une suite de mesures $(P_n)_{n\geq 1}$ est dite relativement compacte si pour toute sous suite $(n') \subset \mathbb{N}$, il existe $(n'') \subset (n')$ telle que $P_{n''}$ converge faiblement vers une mesure de probabilité.

Théorème 8.2.4 (Prohorov)**.** *Soit (P_n) une suite de mesures dans un espace polonais. Alors (P_n) est équitendue si et seulement si (P_n) est relativement compacte.*

Démonstration : Admis. □

Dans ce cours, en général, les processus que l'on considère sont à trajectoires continues et leurs lois sont donc des mesures de probabiliés sur l'espace des fonctions continues $\mathcal{C}^0(T, \mathbb{R})$ polonais pour la topologie uniforme associée à la norme uniforme $||f||_\infty = \sup_{t\in T} |f(t)|$. Dans ce cas, on a un critère d'équitension plus explicite :

Théorème 8.2.5. *Soit (P_n) la suite des lois de processus X_n à trajectoires continues. On suppose qu'il existe $a, b, C > 0$ tels que pour tout $s, t \in [0, 1]$,*

$$\sup_{n\geq 1} \mathbb{E}\big(|X_n(t) - X_n(s)|^a\big) \leq C|t - s|^{1+b}$$

Alors la suite (P_n) est équitendue.

Démonstration : Admis. On notera une ressemblance d'énonciation avec le théorème de Kolmogorov-Centsov. □

Théorème 8.2.6. *Soit $\left(P^{(n)}\right)$ une suite équitendue dans $\mathcal{C}^0(T, \mathbb{R})$ telle que les lois fini-dimensionnelles $P^{(n)}_{t_1,...,t_p}$ converge faiblement pour tout $t_1, ..., t_p \in T$*

$$P^{(n)}_{t_1,...,t_p} \Rightarrow P_{t_1,...,t_p}$$

Alors il existe P de lois fini-dimensionnelles $\left\{P_{t_1,...,t_p} | t_1, ..., t_p \in T,\ p \in \mathbb{N}^*\right\}$ telle que

$$P^{(n)} \Rightarrow P$$

En réalité, la vraie condition supplémentaire nécessaire dans ce théorème est la relative compacité de la suite (P_n) mais d'après le théorème de Prohorov, c'est équivalent à l'équitension de la suite pour laquelle on dispose de critère explicites.

Démonstration : Pour montrer que $P^{(n)} \Rightarrow P$, il suffit de voir que pour toute $(n') \subseteq \mathbb{N}$, il existe $(n'') \subseteq (n')$ telle que $P^{(n'')} \Rightarrow P$.
En effet, on montre qu'alors pour f continue bornée, on a

$$\int f dP^{(n)} \to \int f dP$$

Si ce n'était pas le cas alors il existerait f continue bornée tel que ce n'est pas le cas, i.e. il existerait $\varepsilon > 0$ et $(n') \subseteq (n)$ tels que

$$\left| \int f dP^{(n')} - \int f dP \right| > \varepsilon,\ (*)$$

Or par l'équitension (en fait par relative compacité mais d'après le théorème de Prohorov c'est équivalent), il existe $(n'') \subseteq (n')$ tel que $P^{(n'')} \Rightarrow P$. En particulier, d'après le théorème porte-manteau

$$\int f dP^{(n)} \to \int f dP$$

Comme $\left(\int f dP^{(n')} \right)$ est une suite extraite de $\left(\int f dP \right)$, il y a contradiction avec $(*)$, ce qui justifie la convergence cherchée.
Soit donc $(n') \subseteq (n)$ une sous-suite. Par équitension (relative compacité), il existe $(n'') \subseteq (n')$ et Q tels que $P^{(n'')} \Rightarrow Q$.
En particulier, la convergence des lois fini-dimensionnelles exige que

$$P^{(n'')}_{t_1,...,t_p} \Rightarrow Q_{t_1,...,t_p}$$

Par hypothèse, on a que $P^{(n)}_{t_1,...,t_p} \Rightarrow P_{t_1,...,t_p}$ et donc en particulier $P^{(n'')}_{t_1,...,t_p} \Rightarrow P_{t_1,...,t_p}$. On doit donc avoir $Q_{t_1,...,t_p} = P_{t_1,...,t_p}$ pour tout $t_1, ..., t_p \in T$ et comme la loi est entièrement caractérisée par ses lois fini-dimensionnelles : $P = Q$. Finalement $P^{(n'')} \Rightarrow P = Q$ ce qui conclut par ce que l'on a dit sur la première partie de la preuve. □

8.2.4 Résumé

Pour résumé, on a commencé à travailler sur la loi d'un processus stochastique en énonçant un cadre général

- On considère un processus stochastique $(X_t)_{t \geq 0}$ défini sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ à valeurs dans $\mathbb{R}$.
- Pour chaque $\omega \in \Omega$, on a une trajectoire $t \mapsto X_t(\omega)$.

On a ensuite défini $E = \mathcal{C}^0(\mathbb{R}_+, \mathbb{R})$ (trajectoires continues), muni

- de la topologie de la convergence uniforme sur les compacts.
- de la distance

$$d(x, y) = \sum_{n=1}^{\infty} \frac{\min\{||x - y||_{\infty,n}, 1\}}{2^n}$$

C'est un espace polonais (métrique séparable, complet).

On a ensuite défini la loi du processus X comme la mesure image de $\mathbb{P}$ sur $(\mathcal{C}^0(\mathbb{R}_+, \mathbb{R}), \mathcal{B}(\mathcal{C}^0(\mathbb{R}_+, \mathbb{R})))$. C'est la loi de la trajectoire entière et pas seulement des marginales finies.

On définit ensuite une tribu sur l'espace des trajectoires

$$\sigma(Cyl) = \sigma\{x \mapsto (x(t_1), ..., x(t_n)) | t_1, ...t_n \in \mathbb{R}_+\}$$

Il en suit la proposition clé de cette section :

$$\mathcal{B}(\mathcal{C}^0(\mathbb{R}_+,\mathbb{R})) = \sigma(Cyl)$$

On peut aussi définir la loi du processus par ses projections finies.

La section d'après concerne la régularité des trajectoires et plus precisément sous quelles conditions un processus stochastique admet des versions continues (ou Höldériennes).
On voit alors le théorème de Kolmogorov-Centsov qui sert à nous montrer cette existence de version Höldériennes et aussi à nous garantir que les lois de processus étudiées sont bien portées par $\mathcal{C}^0(\mathbb{R}_+,\mathbb{R})$, donc l'espace où on fera la convergence faible (en loi).

Pour cela, on considère une suite de loi (P_n) sur E défini plus haut.

$$P_n \Rightarrow p \Leftrightarrow \int f dP_n \to \int f dP, \forall f \in \mathcal{C}_b(E)$$

ou encore si $X_n \sim P_n$ et $X \sim P$ alors $X_n \Rightarrow X$.
On énonce ensuite le théorème porte-manteaux qui énonce plusieurs caractérisation équivalente de la convergence faible.

Vient ensuite le tend de la notion de tension

$$\forall \varepsilon > 0,\ \exists K_\varepsilon \text{ compact} \mid \inf_n P_n(K_\varepsilon) > 1 - \varepsilon$$

avec le théorème de Prohorov qui dit que dans un espace polonais, la tension d'une famille de mesure $\{P_n\}$ est équivalente à la compacité séquentielle pour la convergence faible.

Cela s'applique dans notre cas car l'espace $\mathcal{C}(\mathbb{R}_+,\mathbb{R})$ étant polonais, les lois $\mathcal{L}(X_n)$ sont faiblement relativement compact si elles sont tendues.

La tension se vérifie par des conditions sur les trajectoires :

- Bornes en probabilité.
- Régularité Höldérienne en moyenne (Kolmogorov).
- Critères d'Aldous (dans le cas càdlàg).

Comme on parle de compacts, il sera aussi bon d'avoir des notions d'Analyse fonctionnelle et de toologie (par exemple le théorème d'Arzelà-Ascoli, très utilisé dans ce cas pour comprendre la tension des lois de processus à trajectoires continues).

8.3 Processus gaussien

Dans ce chapitre, on commence par présenter la classe des processus gaussiens dont on introduit d'abord la loi, puis la régularité des trajectoires est considéré. On décrit ensuite la notion d'espace gaussien, pour finir par donner des exemples de tels processus.

8.3.1 Lois des processus gaussiens

Définition 8.3.1 (Processus gaussien)**.** Un processus stochastique $(X_t)_{t\in T}$ est gaussien si toutes ses lois fini-dimensionnelles $\mathcal{L}(X_{t_1},...,X_{t_p})$ sont gaussiennes (pour tout $p\in\mathbb{N}^*$ et $t_1,...,t_p\in T$).
Autrement dit $X=(X_t)_{t\in T}$ est gaussien si et seulement si toute combinaison linéaire de ses marginales $a_1X_{t_1}+...+a_pX_{t_p}$ suit une loi gaussienne (pour tout $p\in\mathbb{N}^*$, $t_1,...,t_p\in T$ et $a_1,..,a_p\in\mathbb{R}$).

Il y a alors des conséquences immédiate quand un processus X est gaussien :

- Toutes les marginales d'un processus gaussien sont gaussiennes.
- Toute combinaison linéaire de marginales d'un processus gaussien est encore gaussienne.

Il est connu que la loi d'un vecteur gaussien $(X_{t_1},...,X_{t_p})$ est déterminée (par exemple via sa fonction caractéristique) par le vecteur moyenne

$$m_X=\big(\mathbb{E}(X_{t_1}),...,\mathbb{E}(X_{t_p})\big)$$

et la matrice de covariance

$$\Sigma_X=\big(\mathrm{Cov}(X_{t_i},X_{t_j})_{1\le i,j\le p}\big)$$

On comprend dès lors que toutes les lois fini-dimensionnelles d'un processus gaussien (donc la loi du processus entier) est connue dès qu'on se donne la fonction moyenne $m(t)=\mathbb{E}(X_t)$ et l'opérateur de covariance $K(s,t)=\mathrm{Cov}(X_s,X_t)$.
En effet, la loi fini-dimensionnelle de $(X_{t_1},...,X_{t_p})$ est alors la loi gaussienne $\mathcal{N}(m_p,K_p)$ de dimension p avec

$$m_p=(m(t_1),...,m(t_p)),\ K_p=\big(K(t_i,t_j)\big)_{1\le i,j\le p}$$

Les fonctions m et K définissent donc toutes les lois fini-dimensionnelles de X et donc aussi sa loi en tant que processus. Observons en plus que

- K est symétrique : $K(s,t)=K(t,s)$;
- K est de type positif, i.e. si $c:T\to\mathbb{R}$ est une fonction à support fini alors

$$\sum_{s,t\in T}c(s)c(t)K(s,t)=\mathrm{Var}\bigg(\Big(\sum_{s\in T}c(s)X_s\Big)^2\bigg)\ge 0$$

Maintenant que l'on a cela, il est naturel de se poser la question suivante :
Etant donnés une fonction m sur T et un opérateur K sur $T\times T$, existe-t-il un processu gaussien X admettant m pour fonction moyenne et K pour opérateur de covariance ?

Théorème 8.3.1. *Soit K une fonction symétrique de type positif sur $T\times T$. Il existe alors un processus gaussien (centré) dont la fonction de covariance est K.*

Autrement dit, on veut construire un processus dont les lois fini-dimensionnelles sont gaussiennes centrées et dont $\mathbb{E}(X_sX_t)=K(s,t)$.

Démonstration : Quitte à considérer ensuite le processus $X(t)+m(t)$, on considère le cas simplifié d'un processus centré. Il s'agit alors d'une application simple du théorème d'extension de Kolmogorov.
On construit une probabilité $\mathbb{P}^{(K)}$ sur l'espace mesurable $(\mathbb{R}^T, \sigma(Cyl))$ de telle sorte que sous $\mathbb{P}^{(K)}$ le processus des coordonnées $X_t(\omega) = \omega(t)$ (dit processus canonique) est un processus gaussien de covariance K.
Pour cela, si $\{t_1, ..., t_p\} \subseteq T$, on construit d'abord une probabilité $\mathbb{P}^{(K)}_{\{t_1,...,t_p\}}$ sur $\mathbb{R}^{\{t_1,...,t_p\}} \sim \mathbb{R}^p$ comme la loi du vecteur gaussien $(X_{t_1}, ..., X_{t_p})$ de matrice de covariance $\Sigma_{ij} = \big(K(t_i,t_j)\big)_{1\leq i,j\leq p}$ (qui existe d'après les rappels gaussiens puisque la matrice est symétrique positive). On a donc défini toutes les lois fini-dimensionnelles qu'on veut imposer au processus.
On vérifie aisément que les lois $\mathbb{P}^{(K)}_{\{t_1,...,t_p\}}$ satisfont les propriétés de compatibilité du théorème d'extension de Kolmogorov qui s'applique donc (cela vient du fait que les lois gaussiennes sont stables par marginalisation, la marginale d'un vecteur gaussien reste gaussienne avec la sous-matrice de covariance correspondante).
On peut donc recoller toutes ces lois fini-dimensionnelles pour construire une loi globale $\mathbb{P}^{(K)}$ sur les trajectoires $\omega : T \to \mathbb{R}$.
La loi $\mathbb{P}^{(K)}$ ainsi construite sur $\big(\mathbb{R}^T, \sigma(Cyl)\big)$ est bien celle d'un processus gaussien puisque par construction toutes ses lois fini-dimensionnelles sont gaussiennes avec pour fonction de covariance K. □

Exemple 8.11 (Processus stationnaire). Dans cet exemple on veut construire un processus gaussien stationnaire $(X_t)_{t\in\mathbb{R}}$, i.e.

$$\big(X_{t_1+h}, ..., X_{t_p+h}\big) =^{\mathcal{L}} \big(X_{t_1}, ..., X_{t_p}\big)$$

Autrement dit, sa loi ne dépend pas du temps absolu, mais uniquement des décalages entre les temps.
La clé sera de construire une fonction de covariance stationnaire, i.e.

$$K(s,t) = \mathrm{Cov}(X_s, X_t) = \kappa(t-s)$$

qui ne dépend que de la différence $t-s$.

On considère donc le cas $T = \mathbb{R}$ et on se donne une mesure finie symétrique ($\nu(-A) = \nu(A)$) sur $\mathbb{R}$. On pose alors la représentation spectrale de la covariance (Transformée de Fourier de ν)

$$K(s,t) = \int_{\mathbb{R}} e^{iu(t-s)}\nu(du)$$

On vérifie que K est un opérateur symétrique

$$K(s,t) = \int_{\mathbb{R}} e^{-iu(s-t)}\nu(du) = \int_{\mathbb{R}} e^{-iu(s-t)}\nu(-du) =^{v=-u} \int_{\mathbb{R}} e^{iv(s-t)}\nu(dv) = K(t,s)$$

un opérateur réel

$$K(s,t) = \int_{\mathbb{R}} \cos(u(t-s))du + i \underbrace{\int_{\mathbb{R}} \sin(u(t-s))du}_{=0,\text{ car sin impaire et } \nu(-A)=\nu(A)} = \int_{\mathbb{R}} \cos(u(t-s))du \in \mathbb{R}$$

La symétrie de ν annule la partie imaginaire (les phases opposées se compensent) et ne laisse que le cos, i.e. la partie réelle du noyau harmonique.
De plus, K est de type positif

$$\sum_{s,t\in T} c(s)c(t)K(s,t) = \int \left|\sum_{s\in T} c(s)e^{ius}\right|^2 \nu(du) \geq 0$$

La fonction K possède la propriété supplémentaire de dépendre seulement de la différence $t-s$. On parle de stationnarité faible (ou de deuxième ordre) et on écrit alors $K(t,s) = k(|t-s|)$. On en déduit aussitôt que le processus (centré) X associé à K, par le théorème précédent, est stationnaire (au sens strict), c'est-à-dire pour tout choix de $p \in \mathbb{N}^*$ et $t_1, ..., t_p \geq 0,\ h \in \mathbb{R}$, on a

$$\left(X_{t_1+h}, ..., X_{t_p+h}\right) =^{\mathcal{L}} \left(X_{t_1}, ..., X_{t_p}\right),\ \forall h \in \mathbb{R}$$

En effet, pour un processus gaussien, la loi du vecteur $(X_{t_1}, ..., X_{t_p})$ est entièrement déterminée par son espérance (ici 0) et sa covariance. Ainsi, si K ne dépend que de $t_i - t_j$ alors la stationnarité découle.

Réciproquement, il est aussi vrai que si (X_t) est un processus gaussien stationnaire, continu dans L^2, i.e. $\lim_{s\to t} \mathbb{E}\big((X_t - X_s)^2\big) = 0$, la fonction de covariance de X est de la forme

$$K(s,t) = \int_{\mathbb{R}} e^{iu(t-s)}\nu(du)$$

$$k(\tau) = \int_{\mathbb{R}} e^{iu\tau}\nu(du)$$

C'est le théorème de Bochner : Une fonction $k(\tau)$ est positive définie si et seulement si elle est la transformée de Fourier d'une mesure finie positive. Autrement dit,

$$k(\tau) \text{ est une covariance stationnaire } \Leftrightarrow\ \exists \nu \geq 0,\ k(\tau) = \int e^{iu\tau}\nu(du)$$

La mesure ν est la mesure spectrale du processus, elle joue le rôle de densité d'énergie fréquentielle : elle détermine comment les fréquences u contribuent à la covariance. Elle véhicule beaucoup d'informations concernant le processus.
Par exemple, on a $K(0) = \int e^{iu.0}\nu(du) = \int 1\nu(du) = \mathrm{Var}(X_t) = \nu(\mathbb{R})$. Donc la variance du processus est la masse totale de la mesure spectrale.

De façon générale pour un processus stationnaire, on a

Proposition 8.3.1 (Processus stationnaire)**.** *Soit X un processus stationnaire L^2 centré et de fonction de covariance k. On a équivalence entre*

1. *La fonction k est continue en 0 ;*
2. *Le processus X est L^2-continu*
$$\lim_{s\to t} \mathbb{E}\big((X_t - X_s)^2\big) = 0$$

3. La fonction k est continue partout.

Démonstration : On rappelle que pour un processus stationnaire centré

$$K(t,s) = \mathbb{E}(X_t X_s) = k(t-s)$$

donc la covariance ne dépend que de la différence des temps. En particulier,

$$k(h) = \mathbb{E}(X_{t+h}X_t),\ k(0) = \mathbb{E}(X_t^2)$$

1. $\Leftrightarrow$ **2.** Comme X est centré, on a

$$\mathbb{E}\big((X_s - X_s)^2\big) = \mathbb{E}(X_t^2) + \mathbb{E}(X_s^2) - 2\mathbb{E}(X_t X_s) = K(t,t) + K(s,s) - 2K(t,s)$$

$$= 2k(0) - 2k(|t-s|)$$

En prenant la limite, on comprend l'équivalence via la L^2-continuité. De plus **3.** $\Rightarrow$ **1.** est évident, il ne reste qu'à conclure avec **2.** $\Rightarrow$ **3.**.
On veut montrer que si k est continue en 0 alors k est continue partout. Pour cela, on part de la stationnarité

$$k(t+h) = \mathbb{E}(X_{t+h}X_0)$$

et on cherche à montrer que $k(t+h) \to k(t)$ quand $h \to 0$. Regardons la différence

$$|k(t+h) - k(t)| = |\mathbb{E}(X_{t+h}X_0) - \mathbb{E}(X_t X_0)| = |\mathbb{E}((X_{t+h} - X_t)X_0)|$$

On applique l'inégalité de Cauchy-Schwarz,

$$|\mathbb{E}((X_{t+h} - X_t)X_0)| \leq \sqrt{\mathbb{E}\big((X_{t+h} - X_t)^2\big)}\sqrt{\mathbb{E}(X_0^2)}$$

Cependant, $\mathbb{E}(X_0^2) = k(0)$ et la L^2-continuité donne

$$|k(t+h) - k(t)| \leq \underbrace{\sqrt{\mathbb{E}\big((X_{t+h} - X_t)^2\big)}}_{\to 0}\sqrt{k(0)} \to_{h\to 0} 0$$

Autrement dit, k est continue en tout point t.$\square$

$\to$ Une petite note, dans le cas général d'un processus non-stationnaire, si on considère un processus aléatoire quelconque X, on définit la fonction de covariance comme une fonction de deux variables $K(t,s)$. Dans le cas particulier d'un processus (faiblement) stationnaire $K(t,s) = k(t-s)$, $K(t+h,t) = k(h)$ et k est appelée fonction de covariance stationnaire.

Au passage, on a utilisé que la stricte stationnarité d'un processus gaussien est équivalente à la stationnarité faible.

Proposition 8.3.2. *Un processus gaussien X est stationnaire si et seulement si $t \mapsto \mathbb{E}(X_t)$ est constante et $K(t,s) = k(t-s)$ (on parle de stationnarité faible).*

Démonstration : Il est clair que ces conditions sont nécessaires, que le processus soit gaussien ou pas : comme par stationnarité, on a $\mathcal{L}(X_t) = \mathcal{L}(X_s)$ pour tout t, s, on a $\mathbb{E}(X_s) = \mathbb{E}(X_t)$ et donc la fonction moyenne est constante ; de même $\mathcal{L}(X_t, X_s) = \mathcal{L}(X_{t+h}, X_{s+h})$ pour tout t, s, h alors on a $\mathrm{Cov}(X_t, X_s) = \mathrm{Cov}(X_{t+h}, X_{s+h})$ et donc la covariance ne dépend que de la différence $t - s$.
Elles sont suffisantes seulement dans le cas gaussien puisque dans ce cas, la loi est caractérisée par $t \mapsto \mathbb{E}(X_t)$ et par $K(s,t)$. Si bien qu'une translation dans les paramètres ne modifie pas ces fonctions sous les hypothèses de la proposition et donc, ne modifie pas non plus la loi. □

8.3.2 Régularité gaussienne

Des bonnes conditions pour avoir une version assez régulière d'un processus gaussien sont données dans le résultat suivant, conséquence du théorème de Kolmogorov-Centsov dans le cadre gaussien.

Théorème 8.3.2 (Kolmogorov-Centsov gaussien)**.** *Soit X un processus gaussien centré, de fonction de covariance $K(s,t)$. On suppose qu'il existe $\alpha > 0$ et $0 < C < \infty$ tels que pour tout $s, t > 0$:*

$$K(t,t) + K(s,s) - 2K(s,t) \leq C|t-s|^{\alpha}$$

Alors il existe une version continue $\tilde{X}$ de X. De plus, pour tout $0 < \gamma < \frac{\alpha}{2}$, les trajectoires de $\tilde{X}$ sont presque sûrement höldériennes de coefficient γ.

Démonstration : A partir de

$$\mathbb{E}\big(|X_t - X_s|^2\big) = \mathbb{E}(X_t^2) + \mathbb{E}(X_s^2) - 2\mathbb{E}(X_t X_s) \leq C|t-s|^{\alpha}$$

On ne peut pas appliquer le théorème e Kolmogorov-Centsov car rien ne garanti $\alpha > 1$, alors que cette condition est requise pour l'application du théorème.
On va plutôt s'intéresser à $\mathbb{E}\big(|X_t - X_s|^{2m}\big)$. On rappelle alors que d'après les rappels gaussiens, pour X une variable aléatoire normale centrée, on a :

$$\mathbb{E}(X^{2m}) = \frac{(2m)!}{2^m m!}\mathrm{Var}(X)^m$$

Il vient alors que pour tout $m \geq 1$:

$$\mathbb{E}\big(|X_t - X_s|^{2m}\big) \leq C^m \frac{(2m)!}{2^m m!}|t-s|^{m\alpha}$$

Comme cela est valable pour tout $m \geq 1$, on choisit l'entier m tel que $m\alpha > 1$. D'après le théorème de Kolmogorov-Centsov, avec

$$b = m\alpha - 1,\ a = 2m,\ \text{et}\ \frac{b}{a} = \frac{m\alpha - 1}{2m}$$

Il existe un version $\frac{m\alpha-1}{2m}$-höldérienne de X, pour tout $m < \frac{1}{\alpha}$. Comme

$$\lim_{n\infty} \frac{m\alpha - 1}{2m} = \frac{\alpha}{2}$$

Il existe une version γ-höldérienne de X pour tout $\gamma < \frac{\alpha}{2}$.
Finalement, les versions höldériennes pour des exposants $\gamma \neq \gamma'$ coïncident nécessairement par indistinguabilité. □

8.3.3 Espace gaussien

On rappelle que $L^2(\Omega, \mathcal{F}, \mathbb{P})$ est un espace de Hilbert pour le produit scalaire

$$\langle X, Y \rangle = \mathbb{E}(XY)$$

Définition 8.3.2 (Espace gaussien)**.** Un espace gaussien (centré) est un sous-espace fermé de $L^2(\Omega, \mathcal{F}, \mathbb{P})$ formé de variables gaussiennes centrées.

Par exemple, si $X = (X_1, ..., X_p)$ est un vecteur gaussien centré de $\mathbb{R}^p$, alors $\text{Vect}(X_1, ..., X_p)$ est un espace gaussien (c'est un sous-espace fermé constitué de combinaisons linéaires d'un vecteur gaussien, elles sont donc gaussiennes).

Proposition 8.3.3. *Si $X = (X_t)_{t \in T}$ est un processus gaussien, le sous-espace vectoriel fermé de $L^2(\Omega, \mathcal{F}, \mathbb{P})$ engendré par les variables aléatoires X_t, $t \in T$, est un espace gaussien, appelé espace gaussien engendré par le processus X.*

Démonstration : Le sous-espace vectoriel fermé $\overline{\text{Vect}}^{L^2(\Omega,\mathcal{F},\mathbb{P})}(X_t | t \in T)$ est formé des limites dans $L^2(\Omega, \mathcal{F}, \mathbb{P})$ des combinaisons linéaires finies de marginales X_{t_i} de $(X_t)_{t \in T}$. Ces limites sont gaussiennes car

- Comme (X_t) est gaussien, les combinaisons linéaires le sont aussi ;
- Les limites L^2 de variables gaussiennes sont gaussiennes.

□

Si H est un sous-ensemble de $L^2(\Omega, \mathcal{F}, \mathbb{P})$, on note $\sigma(H)$ la tribu engendrée par les variables aléatoires $Y \in H$.

Théorème 8.3.3. *Soit H un espace gaussien et soit $\{H_i | i \in I\}$ une famille de sous-espaces vectoriels de H. Alors les sous-espaces H_i, $i \in I$, sont orthogonaux dans $L^2(\Omega, \mathcal{F}, \mathbb{P})$ si et seulement si les tribus $\sigma(H_i)$, $i \in I$, sont indépendantes.*

C'est une généralisation des propositions qu'on a pu voir sur les couples et vecteurs gaussiens de dimension $n + p$, concernant l'indépendance et la nullité des covariances. Comme dans ces cas, il est crucial que les espaces H_i soient contenus tous dans un espace gaussien.

Démonstration : Si on suppose que les tribus $\sigma(H_i)$ sont indépendantes, alors pour $i \neq j$ et $X \in H_i$, $Y \in H_j$, on a

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) = 0$$

Ce qui signifie que les espaces H_i sont deux à deux orthogonaux (le produit scalaire étant $\langle X, Y \rangle = \mathbb{E}(XY)$).

Réciproquement, supposons les espaces H_i, $i \in I$, deux à deux orthogonaux. Par définition de l'indépenance d'une famille infinie de tribus, il suffit de démontrer que pour tout indices distincts $i_1, ..., i_p \in I$, les tribus $\sigma(H_{i_1}), ..., \sigma(H_{i_p})$ sont indépendantes. Pour cela, il suffit de montrer que, si $Y_1^1, ..., Y_{n_1}^1 \in H_{i_1}, ..., Y_1^p, ..., Y_{n_p}^p \in H_{i_p}$, les vecteurs $(Y_1^1, ..., Y_{n_1}^p), ..., (Y_1^p, ..., Y_{n_p}^p)$ sont indépendants.

En effet, pour chaque j, les ensembles de la forme $\{Y_1^j \in A_1, ..., Y_{n_j}^j \in A_{n_j}\}$ forment une classe stable par intersection finie qui engendre la tribu $\sigma(H_j)$, et on peut utiliser l'argument classique de classe monotone.
Maintenznt, pour chaque $j \in \{1, ..., p\}$, on considère $Z_1^j, ..., Z_{m_j}^j$ une base orthonormée de $\text{Vect}(Y_1^j, ..., Y_{n_j}^j)$. La matrice de covariance du vecteur

$$(Z_1^1, ..., Z_{m_1}^1, Z_1^2, ..., Z_{m_2}^2, ..., Z_1^p, ..., Z_{m_p}^p)$$

est alors la matrice identité (car pour $i \neq j$, $\mathbb{E}(Z_l^j Z_k^i) = 0$ via l'orthogonalité de H_i et H_j, et pour $i = j$ c'est dû au choix de Z_l^i, $1 \leq l \leq m_i$, base orthonormée de H_i). Ce vecteur est gaussien car ses composantes sont dans H qui est un espace gaussien. Les composantes sont indépendantes, car les covariances sont nulles. On conclut alors que les vecteurs

$$(Z_1^1, ...Z_{m_1}^1), ..., (Z_1^p, ..., Z_{m_p}^p)$$

sont indépendants. Comme pour chaque j le vecteur $(Y_1^j, ..., Y_{n_1}^j)$ est une combinaison linéaire des coordonnées $(Z_1^j, ..., Z_{m_1}^j)$, de manière équivalente les vecteurs

$$(Y_1^1, ..., Y_{n_1}^1), ..., (Y_1^p, ..., Y_{n_p}^p)$$

sont indépendants. □

Corollaire 8.3.1. *Soit H un espace gaussien et K un sous-espace vectoriel fermé de H. On note p_K la projection orthogonale sur K. Soit $X \in H$.*

1. *On a*
$$\mathbb{E}\big(X|\sigma(K)\big) = p_K(X)$$

2. *Soit $\sigma^2 = \mathbb{E}\big((X - p_K(X))^2\big)$. Alors pour tout borélien B de $\mathbb{R}$,*
$$\mathbb{P}\big(X \in B|\sigma(K)\big) = Q(\omega, B)$$
où $Q(\omega, \cdot)$ est la loi $\mathcal{N}(p_K(X)(\omega), \sigma^2)$, i.e.
$$Q(\omega, B) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_B \exp\Big(- \frac{(y - p_K(X)^2)}{2\sigma^2}\Big) dy$$
et avec $Q(\omega, B) = \mathbb{1}_B(p_K(X))$ si $\sigma = 0$.

<u>*Démonstration :*</u> Pour le premier point, soit $Y = X - p_K(X)$ qui est alors orthogonal à K. D'après le théorème précédent, Y est indépendante de $\sigma(K)$. On a donc

$$\mathbb{E}\big(X|\sigma(K)\big) = \mathbb{E}\big(p_K(X)|\sigma(K)\big) + \mathbb{E}\big(Y|\sigma(K)\big) = p_K(X) + \mathbb{E}(Y) = p_K(X)$$

Pour le second point, on écrit pour tout fonction mesurable positive f sur $\mathbb{R}_+$

$$\mathbb{E}\big(f(X)|\sigma(K)\big) = \mathbb{E}\big(f(p_K(X) + Y)|\sigma(K)\big) = \int f\big(p_K(X) + y\big)\mathbb{P}_Y(dy)$$

où $\mathbb{P}_Y$ est la loi de $Y \sim \mathcal{N}(0, \sigma^2)$. On rappelle qu'on a utilisé le fait général suivant : si Z est une variable aléatoire $\mathcal{G}$-mesurable et si $Y \perp\!\!\!\perp \mathcal{G}$ alors

$$\mathbb{E}\big(g(Y,Z)|\mathcal{G}\big) = \int g(y,Z)\mathbb{P}_Y(dy)$$

Le résultat découle aussitôt par loi gaussienne. □

En général, pour une variable aléatoire X dans $L^2(\Omega, \mathcal{F}, \mathbb{P})$ l'espérance conditionnelle est donnée par une projection orthogonale, i.e.

$$\mathbb{E}\big(X|\sigma(K)\big) = p_{L^2(\Omega,\sigma(K),\mathbb{P})}(X)$$

La première assertion du corollaire précedent montre qu'enfait dans le cadre gaussien, la projection orthogonale est à faire directement sur l'espace K, considérablement plus petit que $L^2(\Omega, \sigma(K), \mathbb{P})$. Cette assertion porte aussi le principe de régression linéaire. Par exemple, si (X_1, X_2, X_3) est un vecteur gaussien, la meilleure approximation de X_3 connaissant X_1 et X_2 s'écrit $\lambda_1 X_1 + \lambda_2 X_2$ où λ_1, λ_2 sont déterminés en disant que $X_3 - (\lambda_1 X_1 + \lambda_2 X_2)$ est orthogonal à $\mathrm{Vect}(X_1, X_2)$.

D'une manière plus générale, la loi conditionnelle d'une variable aléatoire réelle X sachant une sous-tribu $\mathcal{G}$ est un noyau $Q(\omega, \cdot)$ $\mathcal{G}$-mesurable, i.e. une application $Q : \Omega \times \mathcal{B}(\mathbb{R}) \to [0,1]$ telle que

- Pour tout ω, $B \mapsto Q(\omega, B)$ est une mesure de probabilité sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.
- Pour tout $B \in \mathcal{B}(\mathbb{R})$, $\omega \mapsto Q(\omega, B)$ est $\mathcal{G}$-mesurable, avec la propriété

$$\mathbb{P}\big(X \in B|\mathcal{G}\big) = Q(\omega, B), \ \forall B \in \mathcal{B}(\mathbb{R})$$

et plus généralement,

$$\mathbb{E}\big(f(X)|\mathcal{G}\big) = \int f(y)Q(\omega, dy)$$

Le deuxième point du corollaire explique alors que dans le cas gaussien, la loi conditionnelle de X sachant la tribu $\sigma(K)$ est explicite, il s'agit de la loi $\mathcal{N}(p_K(X), \sigma^2)$.

8.3.4 Exemples de processus gaussien

Avant d'étudier en détail le mouvement brownien, on décrit briévement ce processus ainsi que quelques processus qui lui sont associés.

Mouvement brownien :

Soit $T = \mathbb{R}_+$, le mouvement brownien (standard) $(B_t)_{t\geq 0}$ est le processus gaussie défini par $\mathbb{E}(B_t) = 0$ et $K(s,t) = \min(s,t)$ (et à trajectoires presque sûrement continues). On l'appelle aussi processus de Wiener (d'où la rencontre avec la notation (W_t) dans la littérature). Noter que $K(t,s) = \min(s,t)$ est bien de type positif puisqu'en écrivant

$$K(t,s) = \int_{\mathbb{R}} \mathbb{1}_{[0,t]}(x)\mathbb{1}_{[0,s]}(x)dx$$

on a

$$\sum_{s,t\in T} c(s)c(t)K(t,s) = \int_{\mathbb{R}} \sum_{s,t\in T} c(s)\mathbb{1}_{[0,s]}(x)c(t)\mathbb{1}_{[0,t]}(x)dx = \int \left(\sum_{t\in T} c(t)\mathbb{1}_{[0,t]}(x)\right)^2 dx \geq 0$$

On peut énoncer quelques propriétés immédiates du mouvement brownien.

- $B_0 = 0$ car la loi de B_0 est $\mathcal{N}(0,0)$ dégénérée en 0.
- (B_t) est à accroissements indépendants. En effet, soit $0 \leq t_1 < t_2 < t_3 < t_4$, on a

$$\begin{aligned}\mathrm{Cov}\big(B_{t_2} - B_{t_1}, B_{t_4} - B_{t_3}\big) &= \mathbb{E}\big((B_{t_2} - B_{t_1})(B_{t_4} - B_{t_3})\big) \\ &= \mathbb{E}\big(B_{t_2}B_{t_4}\big) - \mathbb{E}\big(B_{t_2}B_{t_3}\big) - \mathbb{E}\big(B_{t_1}B_{t_4}\big) + \mathbb{E}\big(B_{t_1}B_{t_3}\big) \\ &= t_2 - t_2 - t_1 + t_1 = 0\end{aligned}$$

  Les variables $B_{t_2} - B_{t_1}$ et $B_{t_4} - B_{t_3}$ sont donc non-corrélées, commes elles sont gaussiennes, elles sont indépendantes. On justifie de même l'indépendance de n accroissements.
- $B_t \sim \mathcal{N}(0,t)$ car $\mathbb{E}(B_t) = 0$ et $\mathrm{Var}(B_t) = K(t,t) = t$.
- Si $s \leq t$, on a $B_t - B_s \sim B_{t-s}$. En effet, par linéarité de l'espérance :

$$\mathbb{E}\big(B_t - B_s\big) = 0$$

  et,

$$\begin{aligned}\mathrm{Var}(B_t - B_s) = \mathrm{Cov}\big(B_t - B_s, B_t - B_s\big) &= \mathrm{Cov}(B_t, B_t) + 2\mathrm{Cov}(B_t, B_s) + \mathrm{Cov}(B_s, B_s) \\ &= t - 2s + s = t - s\end{aligned}$$

  Comme $B_t - B_s$ est de loi normale (combinaison linéaire des marginales d'un processus gaussien), on a

$$B_t - B_s \sim \mathcal{N}(0, t-s) \sim B_{t-s}$$

- Autosimilarité : $\tilde{B}_t = \frac{1}{\sqrt{c}} B_{ct}$, $t \geq 0$, définit encore un mouvement brownien (standard).
- Comportement analogue en 0 et en ∞ : $\tilde{B}_t = tB_{\frac{1}{t}}$, $t \geq 0$ définit encore un mouvement brownien standard. Le processus (B_t) a donc des comportements liés au voisinage de 0 et de ∞.
- Localisation : pour tout $t_0 > 0$, $\tilde{B}_t = B_{t+t_0} + B_{t_0}$, $t \geq 0$ définit encore un mouvement brownien standard. Le processus (B_t) a donc le même comportement en 0 et en tout $t_0 > 0$.
- (B_t) a des trajectoires presque sûrement höldériennes d'ordre γ pour tout $\gamma \in \left]0, \frac{1}{2}\right[$ mais presque sûrement non-dérivables.
  En effet, $\mathbb{E}\big(|B_t - B_s|^2\big) = |t-s|$, donc la continuité höldérienne suit. On admet pour le moment la non-dérivabilité des trajectoires.

Pont Brownien :

Soit $T = [0,1]$, le pont brownien $(B^0_t)_{t\in[0,1]}$ est le processus gaussien centré défini par la fonction de covariance $K(s,t) = \min(s,t) - st$.

Proposition 8.3.4. *On peut définir directement un pont brownien B^0 à partir d'un mouvement brownien B par*

$$B^0_t = B_t - tB_1,\ t \geq 0$$

Démonstration : D'abord, le processus $(B_t - tB_1)_{t\in[0,1]}$ est gaussien, centré puis pour tout $s,t \in [0,1]$ on a

$$\mathrm{Cov}\big(B_t - tB_1, B_s - sB_1\big) = \mathrm{Cov}(B_s,B_s) - t\mathrm{Cov}(B_1,B_s) - t\mathrm{Cov}(B_s,B_1) + ts\mathrm{Cov}(B_1,B_1)$$

$$= \min(t,s) - ts - st + ts = \min(t,s) - ts$$

Il s'agit donc bien d'un pont brownien. □

Réciproquement, on peut construire le mouvement brownien B sur $T = [0,1]$ à partir du pont brownien B^0 et d'une loi normale $N \sim \mathcal{N}(0,1)$ indépendante de B^0 par

$$B_t = B^0_t + tN$$

Le processus $(B^0_t + tN)$ est directement gaussien, centré et le calcul de covariance donne la bonne variance $(\min(t,s))$.

Pour quelques propriétés immédiates,

- $\tilde{B}^0_t = B^0_{1-t}$ définit encore un pont brownien. Le pont brownien est donc symétrique en 0 et en 1 par retournement de temps.
- (B^0_t) a des trajectoires presque sûrement höldériennes d'ordre γ pour tout $\gamma \in \left]0, \frac{1}{2}\right[$ mais presque sûrement non-dérivables. L'argument est le même que pour le mouvement brownien avec $\mathbb{E}\big((B^0_t)^2\big) = t - t^2$.
- Un pont brownien est un mouvement brownien conditionné à valoir 0 à la date $t = 1$ (conditionnement singulier).

Processus d'Ornstein-Uhlenbeck :

8.4 Mouvement brownien

Dans le cadre déterministe, de nombreux phénomènes sont régis par des équations différentielles (ou des EDP). Pour des phénomènes modélisés par un mouvement brownien, on s'attend à avoir des eéquations différentielles faisant intervenir le mouvement brownien. Malheureusement, ce processus a des trajectoires nulles part dérivables et il n'est pas possible de considérer des équations différentielles le faisant vraiment intervenir. Plutôt que de le dériver, on cherchera à intégrer contre

ce processus ce qui permettra de contourner le problème en considérant des équations intégrales (il est d'usage de se ramener à l'écriture symboliques de dérivées et on parlera alors d'équations différentielles stochastiques). Plus tard dans le cours, on définira l'intégrale stochastique pour une large classe de processus (les semi-martingales). Si on se contente du cadre brownien (intégrale et equations différentielles stochastiques pour le mouvement brownien), on parle de calcul d'Itô. Dans ce cadre simplifié, la construction est plus directe.

Dans ce chapitre, nous donnons les principales propriétés (en loi, trajectorielles, variation quadratique), notamment les propriétés de Markov faible et forte. A la fin du chapitre, nous explorons les liens entre le mouvement brownien et l'équation de la chaleur, ce qui correspond à la démarche d'Einstein pour appréhender le mouvement brownien.

8.4.1 Définition, premières propriétés

Le caractère très erratique des trajectoires qui caractérise le mouvement brownien est en général associé à l'observation que le phénomène, bien que très désordonné, présente une certaine homogénéité dans le temps, au sens où la date d'origine des observations n'a pas d'importance.

Définition 8.4.1 (Mouvement brownien)**.** Un mouvement brownien (standard) réel est un processus gaussien centré (B_t) à trajectoires continues de fonction de covariance

$$K(s,t) = \min(s,t)$$

On l'appelle aussi processus de Wiener.

L'opérateur $K(s,t) = \min(s,t)$ est symétrique et de type positif. En effet, si $c : \mathbb{R} \to \mathbb{R}$ est à support borné alors

$$\sum_{s,t\in\mathbb{R}} c(s)c(t)K(s,t) = \sum_{s,t\in\mathbb{R}} c(s)c(t)\min(s,t) = \sum_{s,t\in\mathbb{R}} \int \mathbb{1}_{[0,s]}(x)\mathbb{1}_{[0,t]}(x)dx$$

$$= \int \sum_{s,t} c(s)c(t)\mathbb{1}_{[0,s]}(x)\mathbb{1}_{[0,t]}(x)dx = \int \Big(\sum_{s,t\in\mathbb{R}} c(t)\mathbb{1}_{[0,t]}(x) \Big)^2 dx \geq 0$$

Ainsi, il existe bien un processus gaussien centré de covariance K. Cependant, il n'est pas immédiat que ce processus admette une version à trajectoires continues presque sûrement.

Proposition 8.4.1. *On donne quelques propriétés immédiates du mouvement brownien, certaines ayant déjà été vues dans la partie précédente.*

- $B_0 = 0$;
- $B_t \sim \mathcal{N}(0,t)$;
- (B_t) *est un processus à accroissements indépendants ;*
- *Si* $s \leq t$, *on a* $B_t - B_s \sim B_{t-s}$.

Démonstration : Pour la première, la loi de B_0 est $\mathcal{N}(0,0) = \delta_0$, la loi dégénérée en 0.
Maintenant, $\mathbb{E}(B_t) = 0$ et $\mathrm{Var}(B_t) = K(t,t) = t$, d'où $B_t \mathcal{N}(0,t)$.
Montrons que (B_t) est à accroissements indépendants. Soient $0 \leq t_1 < t_2 < t_3 < t_4$, on a

$$\begin{aligned}\mathrm{Cov}(B_{t_2} - B_{t_1}, B_{t_4} - B_{t_3}) &= \mathbb{E}\big((B_{t_2} - B_{t_1})(B_{t_4} - B_{t_3})\big) \\ &= \mathbb{E}(B_{t_2}B_{t_4}) - \mathbb{E}(B_{t_1}B_{t_4}) - \mathbb{E}(B_{t_2}B_{t_3}) + \mathbb{E}(B_{t_1}B_{t_3})\end{aligned}$$

Cependant, comme (B_t) est centré, $\mathbb{E}(B_t) = 0$ et donc

$$\mathrm{Cov}(B_{t_2}, B_{t_4}) = \mathbb{E}(B_{t_2}B_{t_4}) - \mathbb{E}(B_{t_2})\mathbb{E}(B_{t_4}) = \mathbb{E}(B_{t_2}B_{t_4}) = \min(t_2, t_4)$$

Donc,

$$\mathrm{Cov}(B_{t_2} - B_{t_1}, B_{t_4} - B_{t_3}) = t_2 - t_1 - t_2 + t_1 = 0$$

Si bien que les variables $B_{t_2} - B_{t_1}$ et $B_{t_4} - B_{t_3}$ sont non-corrélées. Comme le vecteur $(B_{t_2} - B_{t_1}, B_{t_4} - B_{t_3})$ est gaussien, les variables sont indépendantes. On généralise de la même manière pour n accroissements.
Maintenant, si $s \leq t$, $\mathbb{E}(B_t - B_s) = 0$ par linéarité, et

$$\mathrm{Var}(B_t - B_s) = \mathrm{Cov}(B_t - B_s, B_t - B_s) = \mathrm{Cov}(B_t, B_t) - 2\mathrm{Cov}(B_t, B_s) + \mathrm{Cov}(B_s, B_s)$$

Si bien que,

$$\mathrm{Var}(B_t - B_s) = t - s$$

Finalement, $B_t - B_s \sim B_{t-s}$.□

On dit enfait que le mouvement brownien est à accroissements indépendants et stationnaires.

Définition 8.4.2 (Définition équivalente)**.** Soit B une famille de variables aléatoires indéxées par le temps. On dit que B est un mouvement brownien si c'est un processus à trajectoires continues tel que

- Pour tout $t \geq 0$, $B_t \sim \mathcal{N}(0,t)$;
- Pour tout $0 \leq t_1 \leq ... \leq t_n$, les variables aléatoires B_{t_1}, $B_{t_2} - B_{t_1}$, ..., $B_{t_n} - B_{t_{n-1}}$ sont indépendants.

Démonstration : On montre ici qu'il s'agit bien d'une équivalence.
On sait déjà, par les propriétés, qu'un mouvement brownien comme défini dans la première définition vérifie les conditions de la deuxième. Donc l'implication directe est immédiate.
Il reste donc à prouver la réciproque.
Soit $s \leq t$, en écrivant $B_t = B_s + B_t - B_s$, par indépendance des accroissements on a

$$\varphi_{B_t} = \varphi_{B_s}\varphi_{B_t - B_s}$$

C'est-à-dire,

$$\varphi_{B_t - B_s}(x) = \varphi_{B_t}(x)\varphi_{B_s}^{-1}(x) = \exp\left(-t\frac{x^2}{2}\right)\exp\left(s\frac{x^2}{2}\right) = \exp\left(-(t-s)\frac{x^2}{2}\right)$$

$$= \varphi_{B_{t-s}}(x)$$

Les accroissements sont donc stationnaires, en particulier ils sont gaussiens. Comme les accroissements sont indépendants, un vecteur d'accroissement $(B_{t_1}, B_{t_2} - B_{t_1}, ..., B_{t_n} - B_{t_{n-1}})$ a pour loi la loi produit de ses marginales qui sont gaussiennes. Un vecteur d'accroissement est donc gaussien. Cependant, comme $(B_{t_1}, ...B_{t_n})$ est une transformation linéaire de ce vecteur d'accroissement, cela reste gaussien. Les lois fini-dimensionnelles (qui caractérisent la loi) étant gaussiennes, alors le processus est gaussien.
Puis, pour $s \leq t$, on a

$$\mathrm{Cov}(B_t, B_s) = \mathbb{E}(B_t B_s) = \mathbb{E}\big((B_t - B_s + B_s)B_s\big) = s$$

Cela confirme donc que le processus défini par la définition ci-dessus est bien le mouvement brownien. □

La probabilité que B_t appartienne à un petit intervalle $[x, x+dx]$ est donc donnée par la densité gaussienne centrée de variance t

$$\mathbb{P}\big(B_t \in [x, x+dx]\big) = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{x^2}{2}\right) dx$$

En particulier, la variable aléatoire B_t qui est une variable aléatoire gaussienne de variance t est comprise entre les nombres $f_1(t) = 2\sqrt{t}$ et $f_2(t) = -2\sqrt{t}$ avec une probabilité de 95%.
On peut montrer que cette propriété est vraie pour toutes les trajectoires brownienne qui est donc comprise entre les deux courbes de f_1 et f_2 avec une probabilité comparable.

En général les phénomènes observés ne sont cependant pas aussi bien normalisés.

Définition 8.4.3 (Mouvement brownien avec drift)**.** On appelle mouvement brownien issu de x, de dérive (ou drift) μ et de coefficient de diffusion σ, le processus

$$X_t = x + \sigma B_t + \mu t$$

Proposition 8.4.2. *Le mouvement brownien (général) X est encore un processus à accroissements indépendants stationnaires et gaussiens. Il est non-centré et tel que $X_0 = x$. De plus, $X_t \sim \mathcal{N}(x + \mu t, \sigma^2 t)$.*

Pour clarifier la chose, quand on parlera de mouvement brownien, il s'agira du mouvement brownien standard B_t sauf mention du contraire.

8.4.2 Propriétés en loi du mouvement brownien

On se fixe un mouvement brownien standard B_t. Systématiquement, pour vérifier qu'on a un mouvement brownien, il s'agit de vérifier qu'on a un processus gaussien, centré, à trajectoires continues et avec la bonne fonction de covariance (c'est souvent là qu'est le challenge).

Symétrie :

Si B est un mouvement brownien, alors $-B$ l'est aussi.
Il suffit de vérifier les propriétés du mouvement brownien. On définit pour tout $t \geq 0$ le nouveau processus $\tilde{B}_t = -B_t$.
$\tilde{B}$ vérifie alors la condition initiale $\tilde{B}_0 = -B_0 = 0$. De plus, (B_t) étant un mouvement brownien, il est à accroissements indépendants et donc $\tilde{B}$ aussi.
Maintenant, pour $s < t$, $\tilde{B}_t - \tilde{B}_s = -(B_t - B_s)$, or $B_t - B_s \sim \mathcal{N}(0, t-s)$ donc $\tilde{B}_t - \tilde{B}_s \sim \mathcal{N}(0, t-s)$.
Finaement, la continuité des trajectoires est évidemment préservée.

Autosimilarité (Propriété d'échelle) :

Pour tout $c > 0$, on pose

$$B_t^{(c)} = \frac{1}{\sqrt{c}} B_{ct}$$

Ce processus défini encore un mouvement brownien (standard). Ce nouveau mouvement brownien a la même loi que (B_t).
En effet, ce nouveau processus est évidemment gaussien (car B l'est). Il est centré, à trajectoires continues. La difficulté réside dans le calcul de la fonction de covariance :

$$\mathrm{Cov}(B_t^{(c)}, B_s^{(c)}) = \mathbb{E}\big(B_t^{(c)} B_s^{(c)}\big) = \frac{1}{c}\mathbb{E}\big(B_{ct} B_{cs}\big) = \frac{\min(ct, cs)}{c} = \min(t, s)$$

Cette propriété montre que c fois B_t se comporte comme un mouvement brownien lu en $c^2 t$: le changement de temps se lit en espace (et réciproquement). En effet, le facteur c agit sur l'échelle de temps et $\sqrt{c}$ sur l'échelle d'espace.

- Si on "accélère" le temps d'un facteur $c > 1$, les fluctuations deviennent $\sqrt{c}$ fois plus grandes.
- Si on "ralentit" le temps $(0 < c < 1)$, les fluctuations deviennent $\sqrt{c}$ fois plus petites.

Autrement dit, les trajectoires du mouvement brownien se ressemblent à toutes les échelles de zoom, c'est une auto-similarité fractale.

Exemple 8.12. Pour $a > 0$, on pose

$$\tau_a = \inf\{t \geq 0 | B_t = a\}$$

On veut exprimer la loi de τ_a en fonction de τ_1 en utilisant la renormalisation $B^{(c)}$.
On choisit $c = a^2$. Alors par définition,

$$B_u^{(a^2)} = \frac{1}{a} B_{a^2 u},\ u \geq 0$$

On effectue ensuite un changement de variables $(t = a^2 u)$ dans le temps d'arrêt τ_a :

$$\tau_a = \inf\{a^2 u | B_{a^2 u} = a\} = a^2 \inf\{u \geq 0 | B_{a^2 u} = a\}$$

Maintenant, on divise

$$B_{a^2 u} = a \Leftrightarrow \frac{1}{a} B_{a^2 u} = 1 \Rightarrow B_u^{(a^2)} = 1$$

Si bien que

$$\tau_a = a^2 \inf\{u \geq 0 | B_u^{a^2)} = 1\} = a^2 \tilde{\tau}_1$$

Maintenant, comme les deux processus ont la même loi, les lois de leur temps de premier passage à 1 coïncident. Donc on a l'égalité en loi

$$\tau_a =^{\mathcal{L}} a^2 \tau_1$$

Autrement dit, le temps nécessaire pour atteindre la hauteur a se comporte comme a^2 fois le temps pour atteindre la hauteur 1.
De cette égalité en loi, on peut en déduire la densité explicite (loi de Lévy) de τ_a. Pour tout $t \geq 0$

$$\mathbb{P}(\tau_a \leq t) = \mathbb{P}\left(\tau_1 \leq \frac{t}{a^2}\right)$$

Le principe de réflexion nous donne la loi de τ_1 et donc on peut en déduire la loi de τ_a.

Inversion du temps :

Le processus $\tilde{B}$ défini par $\tilde{B}_t = tB_{\frac{1}{t}}$ si $t \neq 0$ et $\tilde{B}_0 = 0$ est un mouvement brownien standard. En effet, $\tilde{B}$ est gaussien car ses lois fini-dimensionnelles sont des transformations linéaires de celles de B. Le processus est, de plus, centré et de fonction de covariance

$$\mathrm{Cov}(\tilde{B}_t, \tilde{B}_s) = ts\mathrm{Cov}\big(B_{\frac{1}{t}}, B_{\frac{1}{s}}\big) = ts \min\left(\frac{1}{t}, \frac{1}{s}\right) = \min(t, s)$$

De plus, ses trajectoires sont continues sur $]0, \infty[$ car celles de B le sont sur $\mathbb{R}_+$. Il reste cependant à vérifier que les trajectoires de $\tilde{B}$ sont continues en 0.

$$\mathbb{P}\big(\lim_{t\to 0} \tilde{B}_t = 0\big) = \mathbb{P}\left(\bigcap_{n\geq 1} \bigcup_{p\geq 1} \bigcap_{t\in]0,\frac{1}{p}]\cap\mathbb{Q}} \{|\tilde{B}_t| \leq \frac{1}{n}\}\right) = \mathbb{P}\left(\bigcap_{n\geq 1} \bigcup_{p\geq 1} \bigcap_{t\in]0,\frac{1}{p}]\cap\mathbb{Q}} \{|B_t| \leq \frac{1}{n}\}\right)$$

$$= \mathbb{P}\big(\lim_{t\to 0} B_t = 0\big) = 1$$

Ainsi, le processus B a le même type de comportement en 0 qu'en ∞.

Retournement de temps :

Le processus retourné à l'instant T, $\hat{B}_t^T = B_T - B_{T-t}$ est encore un mouvement brownien sur $[0, T]$. Il est clair que $\hat{B}^T$ est un processus gaussien, centré et à trajectoires continus. De plus, sa fonction de covariance est donnée par

$$\mathrm{Cov}\big(\hat{B}_t^{(T)}, B_s^{(T)}\big) = \mathrm{Cov}(B_T - B_{T-t}, B_T - B_{T-s}) = \min(t, s)$$

Propriété de Markov faible : (ou invariance par translation)

Le mouvement brownien translaté de $t_0 \geq 0$,

$$\overline{B}_t^{(t_0)} = B_{t+t_0} - B_{t_0}$$

est encore un mouvement brownien. De plus, ce processus est indépendant du mouvement brownien arrêté en t_0 : $(B_t)_{0\leq t\leq t_0}$. C'est-à-dire

$$\overline{B}^{(t_0)} \perp\!\!\!\perp \mathcal{F}_{t_0}^B = \sigma(B_s | s \leq t_0)$$

Cette dernière propriété se réécrit dans le cadre classique de la théorie de Markov : indépendance du futur et du passé conditionnellement au présent. En notant $\mathbb{W}_x$ la loi du mouvement brownien issu de $x \in \mathbb{R}$, i.e. de $x + B$ où B est un mouvement brownien habituel, on réécrit

Théorème 8.4.1 (Markov faible)**.** *Soit $t \geq 0$ fixé. Posons $B_s^{'} = B_{t+s}$, $x \in \mathbb{R}$. Alors conditionnellement à $B_t = x$, le processus $B^{'}$ est indépendant de $\mathcal{F}_t^B = \sigma(B_u | u \leq t)$ et a pour loi $\mathbb{W}_x$.*

Démonstration : Pour cela, il suffit de remarquer que $B_s^{'} = \overline{B}_s^{(t)} + B_t$ où $\overline{B}_s^{(t)} = B_{t+s} - B_t$ est un mouvement brownien indépendant de $\mathcal{F}_t^B$ et B_t est une variable $\mathcal{F}_t^B$-mesurable.□

8.4.3 Propriétés trajectorielles du mouvement brownien

Avant de parler de la loi du 0\1 de Blumenthal et autre, on définit d'abord la notion de filtration en temps continu (essentiel pour considérer les martingales au chapitre suivant).

Définition 8.4.4 (Filtration)**.** Une filtration sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ est une famille $(\mathcal{F}_t)_{t\geq 0}$ de sous-tribu telle que pour $s \leq t$, on a $\mathcal{F}_s \subseteq \mathcal{F}_t$.

Si on considère un processus (X_t), on considère souvent la filtration qu'il engendre $\mathcal{F}_t^X = \sigma(X_s | s \leq t)$. Dans ce cas, il est utile d'interpréter une filtration comme une quantité d'information disponible jusqu'à une date donnée : $\mathcal{F}_t^X$ représente ainsi l'information véhiculée par le processus X jusqu'à la date t.
Une filtration est $\mathbb{P}$-complète pour une mesure de probabilité $\mathbb{P}$ si $\mathcal{F}_0$ contient tous les évènements de mesure nulle, i.e. $\mathcal{N} = \{N \in \mathcal{F} | \mathbb{P}(N) = 0\} \subseteq \mathcal{F}_0$. A une filtration $(\mathcal{F}_t)$ on associe

$$\mathcal{F}_{t^+} = \bigcup_{\varepsilon>0} \mathcal{F}_{t+\varepsilon} \text{ et } \mathcal{F}_{t^-} = \bigvee_{\varepsilon>0} \mathcal{F}_{t-\varepsilon}$$

où on rappelle que $\mathcal{A} \vee \mathcal{B} = \sigma(A \cup B)$. La filtration $(\mathcal{F}_t)$ est dite continue à droite (resp. continue à gauche) si pour tout t, on a $\mathcal{F}_t = \mathcal{F}_{t^+}$ (resp. $\mathcal{F}_t = \mathcal{F}_{t^-}$).
Dans la suite, on dira qu'une filtration $\mathcal{F}$ satisfait les conditions habituelles si elle est complète et continue à droite.

Théorème 8.4.2 (Théorème des classes monotones fonctionnel)**.** *Soit E un espace vectoriel fonctionnel monotone (i.e. $f \in E$ est bornée, les constantes sont dans E, si $f_n \in E$ et $f_n \to f$ bornée alors $f \in E$). On suppose que $C \subseteq E$ où C est un ensemble de fonctions stables par multiplication. Alors, E contient toutes les fonctions $\sigma(C)$-mesurables.*

Proposition 8.4.3. *La filtration brownienne $(\mathcal{F}_t^B)$ est continue à droite.*

Démonstration : Pour $t \geq 0$, on va montrer que $\mathcal{F}_{t^+}^{(B)} = \mathcal{F}_t^{(B)}$ en montrant que toute variable aléatoire $\mathcal{F}_{t^+}^{(B)}$-mesurable est $\mathcal{F}_t^{(B)}$-mesurable. Pour cela, on va utiliser le théorème

de classe monotone (fonctionnel).
On considère la tribu

$$\mathcal{G}_t = \sigma(B_{t+s} - B_t | s \geq 0)$$

On commence par prouver que $\mathcal{G}_t$ est indépendante de $\mathcal{F}_{t^+}^{(B)}$. Pour tout $\varepsilon > 0$ par la propriété de Markov (faible), $\mathcal{G}_{t+\varepsilon}$ est indépendante de $\mathcal{F}_{t+\varepsilon}^{(B)}$ et donc indépendante de $\mathcal{F}_{t^+}^{(B)} = \bigcap_{\varepsilon>0} \mathcal{F}_{t+\varepsilon}^{(B)}$. Si bien que

$$\mathcal{G}_{t+\varepsilon} \perp\!\!\!\perp \mathcal{F}_{t^+}^{(B)}$$

Si $t_1 \leq t_2$, on observe que $\mathcal{G}_{t_2} \subseteq \mathcal{G}_{t_1}$ car

$$B_{t_2+s} - B_{t_2} = B_{t_1+(t_2+s-t_1)} - B_{t_1} - \big(B_{t_1+(t_2-t_2)} - B_{t_1}\big)$$

s'exprime en fonction de 2 variables $\mathcal{G}_{t_1}$-mesurables. La famille $\mathcal{G}_{t+\varepsilon}$, $\varepsilon > 0$, est donc croissante quand ε décroit, de limite $\bigvee_{\varepsilon>0} \mathcal{G}_{t+\varepsilon}$.
Par ailleurs, par continuité des trajectoires de B,

$$\forall s \geq 0,\ B_{t+s} - B_t = \lim_{\varepsilon \to 0} (B_{t+s+\varepsilon} - B_t)$$

Ainsi, $B_{t+s} - B_t$ est limite de $\mathcal{G}_{t+\varepsilon}$-mesurable donc aussi mesurable par rapport à $\bigvee \mathcal{G}_{t+\varepsilon}$. Le théorème de classe monotone fonctionnel assure donc que

$$\mathcal{G}_t \perp\!\!\!\perp \mathcal{F}_t^{(B)}$$

Considérons la variable aléatoire positive Y, $\mathcal{G}_t$-mesurable. On va appliquer le théorème de classe monotone fonctionnel avec E l'espace des variables aléatoires bornées $Z \in L^\infty(\mathcal{F})$ qui vérifie

$$\mathbb{E}(ZY) = \mathbb{E}\big(Z\mathbb{E}(Y|\mathcal{F}_t^{(B)})\big)$$

Par le théorème de convergence monotone (avec $Y \geq 0$), on s'assure aisément que E est un espace vectoriel fonctionnel monotoe. Notons

$$\mathcal{M} = \{XZ | X \in L^\infty(\mathcal{F}_t^{(B)}),\ Z \in L^\infty(\mathcal{G}_t)\}$$

On vérifie aisément que $\mathcal{M}$ est une classe multiplicative. De plus, $\mathcal{M} \subseteq E$ car on a

$$\mathbb{E}(XZY) = \mathbb{E}(XY)\mathbb{E}(Z) \text{ (car } \mathcal{F}_t^{(B)} \subseteq \mathcal{F}_{t^+}^{(B)} \perp\!\!\!\perp \mathcal{G}_t)$$

$$= \mathbb{E}\big(X\mathbb{E}(Y|\mathcal{F}_t^{(B)})\big)\mathbb{E}(Z) = \mathbb{E}\big(XZ\mathbb{E}(Y|\mathcal{F}_t^{(B)})\big) \text{ (car } \mathcal{F}_t^{(B)} \perp\!\!\!\perp \mathcal{G}_t)$$

D'après le théorème de classe monotone fonctionnel, l'espace E contient toutes les variables aléatoires bornées et mesurables par rapport à $\mathcal{G}_t \vee \mathcal{F}_t^B = \mathcal{F}^B$ où $\mathcal{F}^B$ est la tribu engendrée par tout le mouvement brownien.
Par conséquent, pour tout $W \in L^\infty(\mathcal{F}^B)$, on a

$$\mathbb{E}(WY) = \mathbb{E}\big(W\mathbb{E}(Y|\mathcal{F}_t^B)\big)$$

En particulier, cela exige $Y = \mathbb{E}(Y|\mathcal{F}_t^B)$ presque sûrement. Comme la tribu est complète, Y coïncide avec une variable $\mathcal{F}_t^B$-mesurable.
Finalement, on peut faire de même pour Y de signe quelconque en écrivant $Y = Y^+ - Y^-$. Ainsi toutes les fonctions bornées $\mathcal{F}_{t^+}^{(B)}$ sont $\mathcal{F}_t^{(B)}$-mesurables et cela prouve $\mathcal{F}_{t^+}^{(B)} = \mathcal{F}_t^{(B)}$.□

En fait, la filtration brownienne est aussi continue à gauche :

$$\mathcal{F}_t^{(B)} = \bigvee_{s<t} \mathcal{F}_s^{(B)}$$

C'est le cas de toute filtration $(\mathcal{F}_t^X)$ engendrée par un processus à trajectoires continues à gauche. En effet, $\mathcal{F}_t^X$ est engendrée par les ensembles $A = \{(X_{t_1}, ..., X_{t_p}) \in \Gamma\}$ avec $0 = t_1 < ... < t_p \leq t$ et $\Gamma \in \mathcal{B}(\mathbb{R}^p)$. Lorsque $t_p < t$ alors $A \in \mathcal{F}_{t_p}^X \subseteq \mathcal{F}_{t^-}^X$. Lorsque $t_p = t$ comme $X_t = \lim_{m\to\infty} X_{s_m}$ pour toute suite $s_m \in [0, t)$ avec $s_m \to t$, on a $A \in \mathcal{F}_{t^-}^X$. Finalement, $\mathcal{F}_{t^-}^X = \mathcal{F}_t^X$.

Une filtration $(\mathcal{F}_{t^+})$ est toujours continue à droite.

Cependant, attention ! Si X est à trajectoires continues $(\mathcal{F}_t^X)$ peut ne pas être être continue à droite, ni $(\mathcal{F}_{t^+}^X)$ à gauche.

Proposition 8.4.4 (Loi du 0\1 de Blumenthal)**.** *La tribu $\mathcal{F}_{0^+}^B$ est triviale, i.e. pour tout $A \in \mathcal{F}_{0^+}^B$, on a $\mathbb{P}(A) = 0$ ou 1.*

Démonstration :

8.4.4 Variation quadratique

8.4.5 Propriété de Markov forte

8.4.6 Equation de la chaleur

8.5 Martingales en temps continu

8.6 Semi-martingales continues

8.7 Intégration stochastique

8.8 Formule d'Itô et conséquences

9 Contrôle Stochastique (M2)

10 Machine Learning (M2)

10.1 Principes du Machine Learning

On va répondre dans cette section, aux questions et principes fondamentaux de l'apprentissage automatique.

10.1.1 Problématique générale de l'apprentissage

Dans cette section, on répond à la question "Qu'est-ce qu'apprendre ?". D'abord d'une manière informelle puis en posant un formalisme mathématique.

L'apprentissage automatique, c'est le processus par lequel une machine améliore ses performances sur une tâche à partir de l'expérience (i.e. des données).
On veut qu'une fonction h soit capable de faire de "bonnes prédictions" sur des exemples qu'elle n'a jamais vus. Par exemple, disons qu'on observe les points $x = 1, 2, 3$ et $y = 2, 4, 6$. On veut apprendre la relation entre x et y, on remarque directement que $y = 2x$ marche sur le jeu d'entraînement. Ce que l'on veut savoir à partir de là c'est, est-ce que la règle $h(x) = 2x$ prévoit bien aussi pour d'autres x jamais observés (comme $x = 4$) ? Si oui, on a appris. Si non, on a juste mémorisé les données.

Posons maintenant un formalisme mathématique pour rendre cela plus rigoureux.

Espaces et inconnues :

- $\mathcal{X}$ est l'espace des entrées (ou des features). Par exemple $\mathcal{X} = \mathbb{R}^d$;
- $\mathcal{Y}$ est l'espace des sorties (labels). Par exemple $\mathcal{Y} = \{0, 1\}$ (classification binaire) ou $\mathbb{R}$ (régression) ;
- D est une distribution sur $\mathcal{X} \times \mathcal{Y}$ qui gouverne la réalité, point faible : elle est inconnue.

En résumé, on a une paire aléatoire

$$(X, Y) \sim D$$

Pour faire écho à ce que l'on disait de façon informelle, en apprentissage supervisé on suppose qu'il existe une fonction inconnue

$$Y = f^*(X) + \varepsilon$$

où $f^*(x)$ est la vraie relation entre X et Y et ε est un bruit aléatoire (moyenne nulle et variance finie), mais on ne connait pas f^*. L'apprentissage consiste alors à approcher f^* par une fonction $h(x)$ qui soit "simple" et qui généralise bien.

Les données d'apprentissage :

On observe un échantillon indépendant et identiquement distribué

$$S = \{(x_1, y_1), ..., (x_n, y_n)\} \sim D^n$$

Autrement dit

$$(x_i, y_i) \sim^{iid} D$$

L'idée c'est que ces données sont un échantillon du monde réel.

L'hypothèse (ou modèle) :

On choisit une famille de fonctions -de prédiction- (ou d'hypothèses) :

$$\mathcal{H} = \{h_\theta : \mathcal{X} \to \mathcal{Y} | \theta \in \Theta\}$$

où Θ est l'espace des paramètres (indexe toutes les hypothèses possibles).

Par exemple,

- $h_\theta(x) = \theta^T x$ pour la régression linéaire ;
- $h_\theta(x) = W_2\sigma(W_1 x + b_1) + b_2$ pour un réseau de neurones. W_1, W_2 sont des matrices de poids, b_1, b_2 des biais et σ est la fonction d'activation (ex. ReLU, sigmoïde... principalement là pour introduire de la non-linéarité). Ici, $\Theta = \{(W_1, b_1, W_2, b_2)\}$ est l'ensemble de tous les paramètres du réseau.

Loss function :

C'est la pénalité pour une mauvaise prédiction

$$l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$$

Par exemple,

- Classification binaire : $l(\hat{y}, y) = \mathbb{1}_{\hat{y} \neq y}$;
- Régression : $l(\hat{y}, y) = (\hat{y} - y)^2$.

Notre loss function tient enfait le rôle de mesure de la qualité.

True risk :

Le "vrai" objectif qu'on aimerait minimiser est le risque attendu :

$$R_D(h) = \mathbb{E}_{(X,Y)\sim D}\big(l(h(X), Y)\big)$$

Intuitivement : "Quelle est la perte moyenne que je ferai sur de nouvelles données tirées de la vraie distribution D ?". Le problème c'est qu'on ne connait pas D donc on ne peut pas calculer $R_D(h)$ directement.

Ainsi, plus formellement l'objectif serait de trouver une hypothèse $\hat{h} \in \mathcal{H}$ qui minimise R_D.

Le risque empirique :

On estime alors empiriquement le true risk à partir de notre échantillon S

$$\widehat{R}_n(h) = \frac{1}{n}\sum_{i=1}^{n} l\big(h(x_i), y_i\big)$$

qui est la moyenne de la perte sur les données observées.

Le problème revient alors à résoudre

$$\widehat{h} = \arg\min_{h\in\mathcal{H}} \widehat{R}_n(h)$$

Risque bayésien :

S'il existait une fonction idéale h^* minimisant le risque sur toutes les fonctions possibles

$$h^* = \arg\min_{h\in\mathcal{H}} R_D(h)$$

On l'appelle classificateur de Bayes. Cependant comme on ne connait pas D on ne peut que s'en approcher par $\hat{h}$.

Règle d'apprentissage :

On définit un algorithme d'apprentissage A :

$$\begin{array}{rccc} A & : & (\mathcal{X}\times\mathcal{Y})^n & \to & \mathcal{H} \\ & & S & \mapsto & A(S) = \hat{h} \end{array}$$

Autrement dit, à chaque échantillon S, l'algorithme renvoie un modèle $\hat{h}$. Le plus souvent, A est basé sur la minimisation du risque empirique

$$\widehat{h} = \arg\min_{h\in\mathcal{H}} \widehat{R}_n(h)$$

C'est ce qu'on appelle l'ERM (Empirical Risk Minimization).

**

Le cadre que nous venons de développer permet de poser les bonnes questions mathématiques :

1. Est-ce que $\hat{h}$ se rapproche de h^*, le meilleur modèle en espérance

$$h^* = \arg\min_{h\in\mathcal{H}} R_D(h)$$

2. Combien faut-il de données n pour avoir une garantie précise sur la différence

$$|R_D(\hat{h}) - R_D(h^*)|$$

3. Quels types de classes $\mathcal{H}$ sont "apprenables" ?

Après tout ça, on a enfin une idée de ce que c'est "apprendre".
C'est comprendre quand et pourquoi l'approximation empirique $\widehat{R}_n(h)$ est proche du risque réel $R_D(h)$.
Autrement dit

$$\text{Apprendre} \iff R_D(\hat{h}) \approx R_D(h^*)$$

Synthèse géométrique :

On peut imaginer que Θ est un espace de dimension p (paramètres) et qu'à chaque point $\theta \in \Theta$ correspond une fonction h_θ. Le risque empirique

$$\widehat{R}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} l\big(h_\theta(x_i), y_i\big)$$

définit une surface sur Θ. L'apprentissage consiste à trouver un minimum global (s'il existe) de cette surface

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \widehat{R}_n(\theta)$$

On peut passer du cadre $\mathcal{H}$ (théorie d'apprentissage abstraite, PAC) au cadre paramétrique Θ (optimisation, Deep Learning). les deux se rejoignent via le prinipe d'ERM.

10.1.2 Risque attendu et risque empirique

On a toujours

$$R_D(h) = \mathbb{E}_{(X,Y)\sim D}\big(l(h(X), Y)\big)$$

C'est la perte moyenne sur des données nouvelles provenant de la distribution vraie.

Concernant le risque empirique

$$\widehat{R}_n(h) = \frac{1}{n}\sum_{i=1}^{n} l(h(x_i), y_i)$$

c'est la perte moyenne sur les données d'entrainement.

10.1.3 Risque empirique et généralisation

10.1.4 Dimension et biais-variance

10.1.5 Introduction au cadre PAC et notion de complexité

10.1.6 Annexes

10.2 Régression linéaire

10.2.1 Formulation du problème

Pour une première modélisation, on peut imaginer une exemple : on cherche à savoir s'il est possible d'expliquer le taux maximal d'ozone de la journée par la température T_{12} à midi en prenant des mesures.
D'un point de vue pratique, le but de cette régression est double :

- Ajuster un modèle pour expliquer O_3 en fonction de T_{12} ;
- Prédire les valeurs d'O_3 pour de nouvelles valeurs de T_{12}.

Pour analyser la relation entre les x_i (température - input) et les y_i (ozone - output/target), on peut qualitativement commencer par plot les données. Puis formellement, on cherche une fonction f telle que

$$y_i \sim f(x_i)$$

Pour préciser le sens de $\sim$ dans ce contexte, il faut se donner un critère quantifiant la qualité de l'ajustement de la fonction f aux données. Il conviendra aussi de se donner une classe de fonction $\mathcal{F}$ dans laquelle est supposée vivre la vraie fonction inconnue.
On l'a vu dans le section précédente, le problème mathématique peut alors s'écrire de la façon suivante

$$\arg\min_{f\in\mathcal{F}} \sum_{k=1}^{n} L(y_i - f(x_i))$$

10.2.2 Régression linéaire simple - OLS

Une première approche naturelle est de considérer que la variable à expliquer y est une fonction affine de la variable explicative x. C'est exactement le principe de la régression linéaire simple. On suppose dans la suite disposer d'un échantillon de n points (x_i, y_i).

Définition 10.2.1 (Modèle de régression linéaire simple)**.** Un modèle de régression linéaire simple est défini par une équation de la forme

$$\forall i \in \{1, ..., n\},\ y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

Ce modèle est dit de Gauss-Markov.

Les quantités ε_i, que l'on appelle erreurs (ou résidus, ou bruits blanc), viennent du fait que les points ne sont jamais parfaitement alignés sur une droite. Ces bruits sont supposés aléatoires et on verra plus tard le cas particulier du bruit gaussien.

Définition 10.2.2 (Hypothèses de la régression linéaire simple)**.** Pour dire des choses sur ce modèle, il nous faut certaines hypothèses concernant le bruit :

1. Centrage : $\forall i,\ \mathbb{E}(\varepsilon_i) = 0$;
2. $\forall (i,j),\ \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \delta_{ij}\sigma^2$.

Il note directement que l'hypothèse 2. est enfait la combinaison de 2 hypothèses qu'il peut être bon d'avoir en tête :

- Non-corrélation : $\forall i \neq j,\ \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$;
- Homoscédasticité (même variance) : $\forall i,\ \mathrm{Cov}(\varepsilon_i, \varepsilon_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2$.

Il est aussi bon de directement remarquer que le modèle de régression peut s'écrire sous forme matricielle

$$Y = \beta_1 \mathbb{1} + \beta_2 X + \varepsilon$$

avec

- $Y \in \mathbb{R}^n$;
- $\mathbb{1} = (1, ..., 1)^T$;
- $X = (x_1, ..., x_n)^T$;
- $\beta_1, \beta_2 \in \mathbb{R}$.

Cette notation est utile pour l'interprétation géométrique du modèle.

Donc maintenant revenons à notre problème après avoir modéliser la situation. Les (x_i, y_i) sont donnés et on veut trouver une fonction affine f qui minimise la quantité

$$\sum_{i=1}^{n} L(y_i - f(x_i))$$

Pour cela, il nous faut préciser la fonction de coût. Deux fonctions sont classiquement utilisées :

- Coût absolu : $L(u) = |u|$;
- Coût quadratique : $L(u) = u^2$.

En pratique, on utilisera la fonction de coût quadratique pour plusieurs raisons :

Premièrement, le problème

$$\min_{\beta} = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

est strictement convexe et donc admet un minimum global/solution analytique. Comparativement au problème

$$\min_{\beta} = \sum_{i=1}^{n} |y_i - x_i^T \beta|$$

nous donne un problème non-différentiable beaucoup plus complexe (pas de gradient classique, pas de solution en forme fermée, temps de calcul plus élevé...).
En effet, la fonction de loss quadratique est différentiable ($\mathcal{C}^\infty$) et de gradient

$$\nabla_\beta \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 = -2X^T (y - X\beta)$$

On a une descente de gradient rapide et stable.
Alors que si on prend la fonction de loss absolue, $\frac{d}{d\varepsilon}|\varepsilon|$ est entre -1 et 1 suivant le signe de ε. Cela

impose des méthodes spéciales et une convergence plus délicate.

On a aussi discuter du cas particulier du bruit gaussien. On verra que si on suppose $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ alors l'estimateur OLS est égale à l'estimateur du maximum de vraisemblance. Cela s'avère très pratique car en pratique beaucoup de phénomènes ont une composante gaussienne.

En résumé, le choix de la fonction de perte quadratique est naturelle dans notre cadre car elle amène une solution fermée "facile", efficace et statistiquement optimale sous bruit gaussien.

Dans ce cadre, on parle de méthode d'estimation par moindres carrés ou OLS (Ordinary Least Squares).

Définition 10.2.3 (Estimateurs des moindres carrés)**.** On appelle estimateurs des moindres carrés ordinaires $\widehat{\beta}_1$ et $\widehat{\beta}_2$ les valeurs minimisant la quantité

$$S(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2$$

C'est-à-dire, la droite des moindres carrés minimise la somme des carrés des distances verticales des points (x_i, y_i) du nuage à la droite ajustée $y_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_i$. On remarque d'ailleurs que ce qu'il y a à l'intérieur du carré, c'est enfait notre bruit donc cela fait du sens qualitativement.

On a donc vu que ce problème était convexe et donc ne posait aucun problème. En effet, S est strictement convexe et admet donc un minimum global en un unique point $(\widehat{\beta}_1, \widehat{\beta}_2)$.

Proposition 10.2.1 (Estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$)**.** *Les estimateurs des Moindrs Carrés Ordinaires ont pour expressions*

$$\hat{\beta}_1 = \overline{y} - \hat{\beta}_2 \overline{x}$$

où

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Démonstration : Comme dit précédemment, la fonction S ci-dessus est strictement convexe en tant que quadratique, elle admet donc un minimum en un point unique $(\widehat{\beta}_1, \widehat{\beta}_2)$ qui est déterminé en annulant les dérivées partielles de S :

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0$$

$$\frac{\partial S}{\partial \beta_2} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0$$

La première équation sur la dérivée partielle donne

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \Longleftrightarrow \sum_{i=1}^{n} y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} x_i \Longleftrightarrow \hat{\beta}_1 = \overline{y} - \hat{\beta}_2 \overline{x}$$

La deuxième équation sur la dérivée partielle donne

$$\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \iff \sum_{i=1}^{n} x_i y_i = \hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2$$

$$\iff \sum_{i=1}^{n} x_i y_i = (\overline{y} - \hat{\beta}_2 \overline{x}) \sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2$$

$$\iff \hat{\beta}_2 =$$

10.2.3 Régression linéaire multiple - OLS

10.2.4 LASSO - Pénalisation L^1

10.2.5 Ridge - Pénalisation L^2

10.2.6 ElasticNet

10.3 Régression logistique

10.4 Principal Component Analysis

10.5 Linear Discriminant Analysis

10.6 Quadratic Discriminant Analysis

10.7 Decision Tree

10.8 Random Forest

10.9 Adaboost

10.10 Gradient Boosting

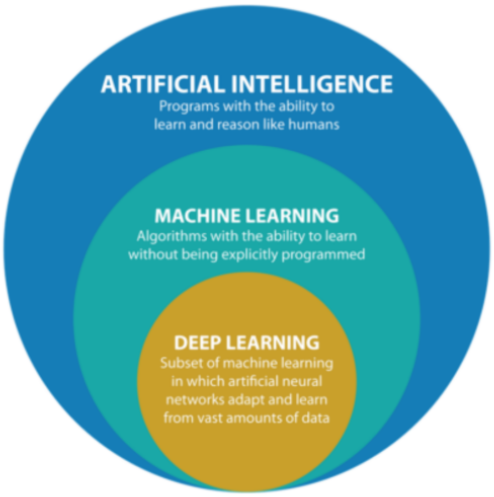
10.11 XGBoost

11 Deep Learning (M2)

11.1 Introduction

In this course, we will discuss :

- Machine Learning
- Neural Netwrks
- Convolutional Neural Networks
- Recurrent Neural Networks
- Generative Adversarial Networks (GAN)
- Applications of Deep Learning to Ficance



Deep neural network or Deep Learning have revolutionized many domains :

- Achieving superhuman performances in image recognition, natural language processing, board and video games, protein folding...
- Close-to-human performances in video processing and self-driving cars...

We easily understand that the potential for Deep Learning to transform financial systems is significant, with successful applications in

- Sentiment analysis
- Credit default detection
- Satellite image analysis

- Portfolio design

Deep Learning is considered a major new trend in the evolution of Artificial Intelligence in recent years for two main reasons :

- Deep Learning can leverage "Big Data" with significantly higher accuracy compared to other Machine Learning methods, especially on image datasets.
- The advancements in hardware (e.g. the release of NVIDIA's GTX 10 series GPUs in 2016 with high computational performance and affordable pricing) made Deep learning research more accessible. It was no longer limited to expensive research labs or large corporations.

11.2 Machine Learning

The fundamental paradigms of Machine Learning include :

- **Supervised (statistical) learning :** The task is to learn a function that maps an input to an output based on example intput-output pairs (i.e. providing output labels).
  Example : Regression problem. Predict Y (wage) based on X_1 (education), X_2 (gender),...
  Example : Classification problem. Predict ups/downs of the S&P500 ($Y \in \{0, 1\}$) based on some macroeconomic variable X_1,X_2,...
- **Reinforcement (statistical) learning :** The task is to learn a mapping function but without explicit input-output examples. Instead, it is based on a reward function tht evaluates the performance of the current state and action.
  Example : Create chatbots to learn how to respond to users. Basic reinforcement learning is modeled as a Markov Decision Process.
- **Unsupervised (statistical) learning :** The task is to learn relationships and structure from data where ther are inputs but no supervised output labels or reward function.
  Example : Clustering problem. Group the customers based on inputs X_1 (age), X_2 (income),...

11.2.1 Black-Box Modelling

Denoting the input or p features as $X = (X_1, ..., X_p)$ and the output as Y, we conjecture that reality can be modeled as the noisy relationship

$$Y = f(X) + \varepsilon$$

where f is some unknown function, to be learned, that maps X into Y, and ε denotes the random noise. This modeling via the function f is what is usually referred to as "Black box", in the sense that no attempt is made at understanding the mapping f, it is simply learned. One of the main reasons to estimate or learn the function f is for prediction, also termed forecasting if the prediction is for a future time : assuming a given input X is available, the output Y can be predicted as

$$\widehat{Y} = \widehat{f}(X)$$

Depending on the nature of the output Y, Machine Learning systems are classified into

- Regression : The output is a real-value number,

- Classification : The output can only take discrete values, such as $\{0, 1\}$ or $\{cat, dog\}$.

In order to learn the function f, we need to be able to evaluate how good a candidate model is. This s conveniently done by defining an error function, also known as a loss (or cost) function or a metric for evaluating model performance. Any appropriate error function for the problem at hand can be used ; the two typical choices for error functions are

- Means Squared Error (MSE) for Regression :

$$\mathbb{E}\big((Y - \widehat{Y})^2\big)$$

- Accuracy (ACC) for Classification :

$$\mathbb{E}\big(\mathbb{1}_{Y=\widehat{Y}}\big)$$

The expectation operator $\mathbb{E}(\cdot)$ is over the distribution of random inputoutput pairs (X, Y). In practice it has to be approximated via the sample mean overs ome observations.

Now, let's talk about underfitting and overfitting.

- The over-simplistic model leads to underfitting
- The best fit model have the training error far from the true error since the learning process can lead to a small training error that is not representative of th actual error, a phenomenon termed overfitting.

To properly evaluate the performance of a system, new data that has not been used during the learning process has to be employed for validation, termed validation data, producing the validation performance, such as the validation MSE or ACC, representative of the true performance.
The divergence between the training MSE and the validation MSE afer some point of complexity implies overfitting : the system is fitting the noise in the training data that is not representative of the validation data.
After that, for testing the chosen model, a new dataset called test data is used.

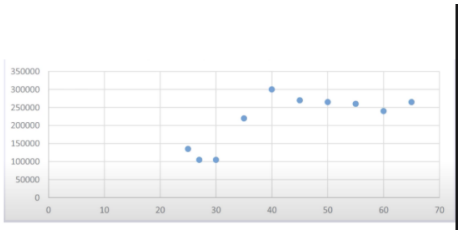
Exemple 11.1 (Regression problem). We want to predict salaries of people from their age. The size of the data sample is $n = 30$. We divide the data sample into

- Training set : $n = 10$
- Validation set : $n = 10$
- Test set : $n = 10$

Candidate model :

1. Linear model : $Y = a(\text{intercept}) + b_1(\text{slope})X$
2. Polynomial of order 2 : $Y = a + b_1X + b_2X^2$
3. Polynomial of order 5 : $Y = a + b_1X + ... + b_5X^5$

Age	Salary
25	135,000
55	260,000
27	105,000
35	220,000
60	240,000
65	265,000
45	270,000
40	300,000
50	265,000
30	105,000



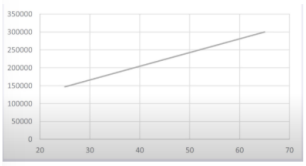


Figure: Linear model

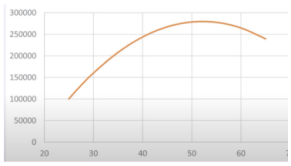


Figure: Polynomial of order 2

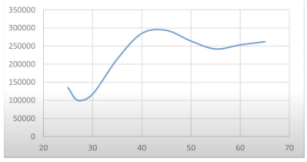


Figure: Best fit model : Polynomial of order 5

	Linear	Polynomial of order 2	Polynomial of order 5
Training set	49,731	32,932	12,902
Validation set	49,990	33,554	38,794
Difference	259	622	25,892
Overfitting ?	No	No	Yes

Table: RMSE of the candidate models and insights

The linear model appears to be over-simplistic, it could suffer from underfitting. The polynomial of order 5 is overfitting based on te validation RMSE. Polynomial of order 2 produces the best fit without overfitting the data.
Now we can ask ourselves that question : How accurate is the chosen model ?
The answer lies in the test set. The accuracy is measured by RMSE on the test set (not training or validation sets).

Can we only use training set and test set ? The answer is : it depends.
Pros of using only training and testing sets :

- Simplicity : Fewer datasets to manage can simplify the workflow.
- Direct evaluation : You can directly assess model performance on unseen data with the testing set.

Cons of not having validation set :

- Hyperparameter tuning : Without a validation set, it becomes challenging to tune hyperparameters effectively. You would need to rely on techniques like cross-validation or use the testing set for tuning, which can lead to overfitting.
- Model selection : You won't have a separate set to evaluate different models or configurations, making it harder to choose the best-performing model.
- Overfitting risk : If you evalute your model on the testing set during training (e.g. tuning based on test performance), you risk overfitting to that specific dataset, which can lead to poor generalization to new data.

Feature	Training set	Validation set	Test set
Purpose	Model learning	Model tuning	Model evaluation
Used in	Model training phase	Model validation phase	Final testing phase
Exposure to model	Directly used	Indirectly used (for tuning)	Never used during training or tuning
Risk of overfitting	High if too small or overused	Medium	Low (if unused during training)

Exemple 11.2 (Classification problem). We want to predict the probability of credit card defauts based on annual income and monthly credit card balance.

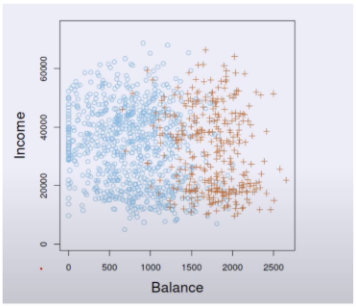


Figure: Defauts in orange, non-defauts in blue

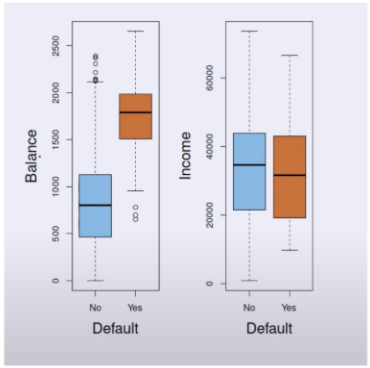
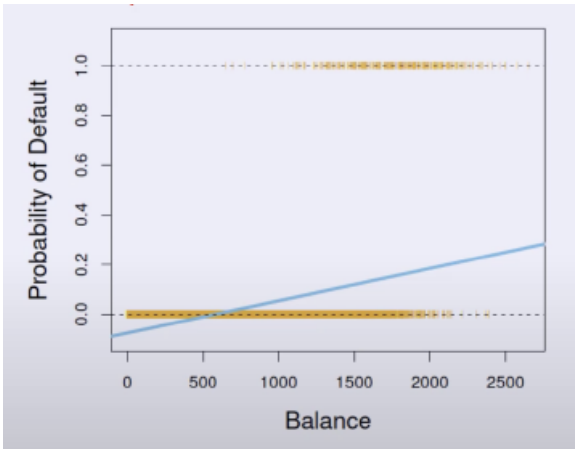


Figure: Boxplot of the data

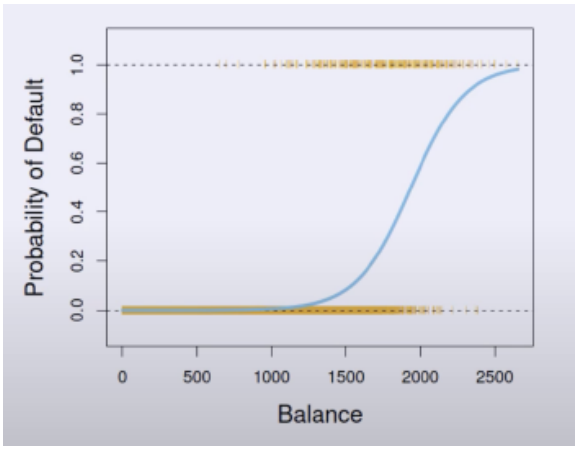
Default : Yes=1, No=0.
One solution is to fit a linear model $\mathbb{P}(Default = 1|Balance) = a + b$.Balance to the data. However, this model produces the negative probabilities which is not good !



Better solution is to fit a logistic model

$$\log\left(\frac{\mathbb{P}(Default = 1|Balance)}{1 - \mathbb{P}(Default = 1|Balance)}\right) = a + b$$

Balance to the data.



11.2.2 Learning the Model

The black-box model f is learned based on the training data by minimizing the training error or optimizing a performance measure.The specific mechanism by which f is learned depends on the particular black-box model. For example, in artificial neural networks, the training algorithms are variations of stochastic gradient descent.
Supervised learning is used to learn the function f based on input/output pairs (i.e., with labels) in order to minimize the error (e.g., MSE or ACC). When explicit labels are not available but the performance of the model can be measured, reinforcement learning can be effectively used.

In practice, overfitting happens when there is too little training data or when the number of parameters (i.e., degrees of freedom) that characterize f is too large, as illustrated in the above

example. To avoid overfitting (to avoid fitting the noise in the training data that is not representative of the rest of the data), two main philosophies have been developed in order to choose an adequate complexity for the model (i.e., the number of degrees of freedom or parameters), termed model assessment (Hastie et al.,2009, Chapter 7) :

- Empirical cross-validation methods : These simply rely on assessing the performance of a learned model f (estimated from training data) on new data termed cross-validation data (note that, once the final model for f has been made, the final performance will be assessed on yet new data termed test data).
- Statistical penalty methods : To avoid reserving precious data for cross-validation, these methods rely on mathematically derived penalty terms on the degrees of freedom ; for example, the Bayesian information criterion (BIC),the minimum description length (MDL), and the Akaike information criterion (AIC).

In practice, we constrain the search to some class of functions f and employ some finite-dimensional parameters to conveniently characterize f. For example, linear models or other classes of nonlinear models can adopt more complicated structures, but always with a finite number of parameters. Over the decades, since the advent of linear regression methods in the 1970s, a plethora of classes of functions have been proposed. The "no free lunch theorem" in statistics : no one method dominates all others overall possible datasets. On a particular dataset, one specific method may work best, but some other method may work better on a different data set. Hence it is an important task to decide which method produces the best results for any given set of data. Selecting the best approach can be one of the most challenging parts of performing statistical learning in practice.

Some machine learning methods that have enjoyed success in ML include (Bishop, 2006 ; Hastie et al.,2009 ; Shalev-Shwartz et Ben-David, 2014 ; Vapnik, 1999) :

- Linear models
- Sparse linear models
- Decision trees
- k-nearest neighbors
- Bagging
- Boosting
- Random Forest (Boosting applied to decision trees)
- Support vector machines (SVM)
- Neural networks, which are the foundation of Deep Learning

Interestingly, some of the more complex models, such as random forests and neural networks, are so-called universal function approximators, meaning that they are capable of approximating any nonlinear smooth function to any desired accuracy, provided that enough parameters are incorporated.

11.2.3 Applications of ML in Finance

ML can be used in finance in a multitude of ways. Here, we discuss the two most obvious applications : time series forecasting and portfolio design. However, there are many other aspects where ML can be used, for example, credit risk (Atiya, 2001), sentiment analysis, outlier detection, asset pricing, bet sizing, feature importance, order market execution, big data analysis (Lopez de Prado, 2019), and so on.

Loopez de Prado (2018b) gives a comprehensive treatment of recent machine learning advances in finance, with extensive treatment of the preprocessing and parsing of data from its unstructured form to the appropriate form for standard ML methods to be applied. On a more practical aspect, reasons why most machine learning funds fail are presented in Lopez de Prado (2018a).

In the realm of time series analysis, there exists a large number of publications addressing different aspects. An overview of machine learning techniques for time series forecasting is provided in Ahmed et al. (2010) and Bontempi et al. (2012), while a comparison between support vector machines and neural networks in financial time series is performed in Cao and Tay (2003). Pattern recognition in time series is covered in Esling and Agon (2012) and a comparison of various classifiers for predicting stock market price direction is provided in Ballings et al. (2015).

11.2.4 Exercises

11.3 Neural Networks

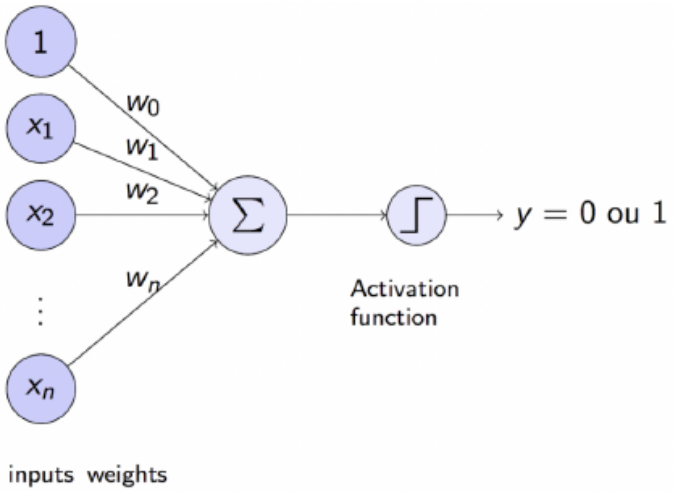
The term Neural Network encompasses a large class of models. First, we focus on "vanilla" NNs which are often referred to as single-layer perceptron or single hidden layer back-propagation networks. We also then cover deeper networks (multilayer perceptron) as well as convolutional NNs.

Neural Networks are not new :

- 1940s-1960s : DL known as cybernetics.
- 1980s-1990s : DL known as connectionism.
- 2006-present : DL.

Since the amount of training data is increasing, NNs have become more useful. Overtime, models have grown in size as both hardware and software have improved. NNs have helped to solve increasingly complicated applications with increasing accuracy.

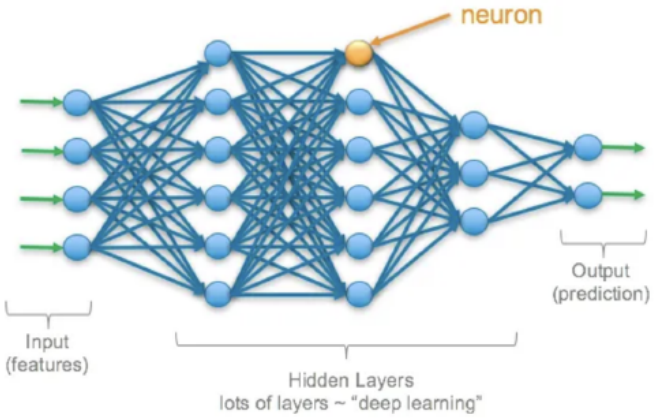
In 1957, Frank Rosenblatt invents the perceptron, a supervised learning algorithm for binary classifiers :



If $\sum_{i=0}^{n} w_i \times x_i < 0$ then $y = 0$, else $y = 1$.

In 1986, the process of back-propagation is described by David E. Rumelhart, Geoffrey E. Hinton and Ronald J. Williams.

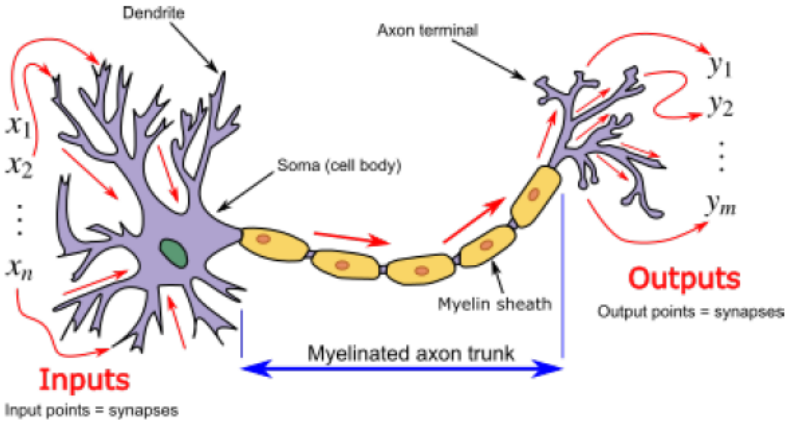
In the following graphic, each circular node represents an artificial neuron and an arrow represents a connection from the output of one artificial neuron to the input of another.



And then of course, DL became more powerful and useful :

- 1997 : IBM Deep Blue Beats Kasparov.
- 2012 : AlexNet learns to recognize images on ImageNet.
- 2015 : AlphaGo beat a human professional Go player.
- 2017 : DeepStack wins professional poker tournament.
- 2019 : Alphastar.
- 2021 : DALL-E and Text A.I.

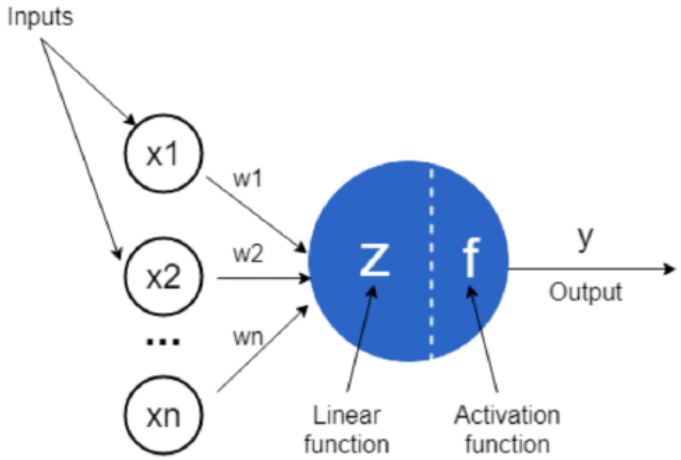
An artificial neural network (ANN) is an interconnected group of nodes, inspired by a simplification of neurons in a brain. It has at least two layers (input and output).



ANNs have the ability to learn and model non-linearities and complex relationships. This is achieved by neurons being connected in various patterns, allowing the output of some neurons to become the input of others. The network forms a directed, weighted graph.

11.3.1 Perceptron

The perceptron is a network with only two layers.



The training data-set consists of N pairs : $\{(X_1, Y_1), ..., (X_i, Y_i), ..., (X_N, Y_N)\}$ where

- X_i is the n-dimensional input vector ;
- $X_{i,j}$ is the j-th element of the vector ;
- $X_{i,0}$ is considered to be 1 ;
- Y_i is the target value (0 or 1) for that input.

Let w_j is the weight of the linear regression over the j-th element. Since it is an iterative algorithm, $w_j(t)$ symbolizes the value of the weights at iteration t. Initialise $w(0)$ to some values. Finally, η is the learning rate, be a small positive number (small steps lessen the possibility of destroying correct classifications).

How to construct a perceptron classifier :

- Select random sample from training set as input.
- Calculate the output :

$$\widehat{Y}_i(t) = f\big(\underbrace{w(t)^T X_i}_{Z_i}\big)$$

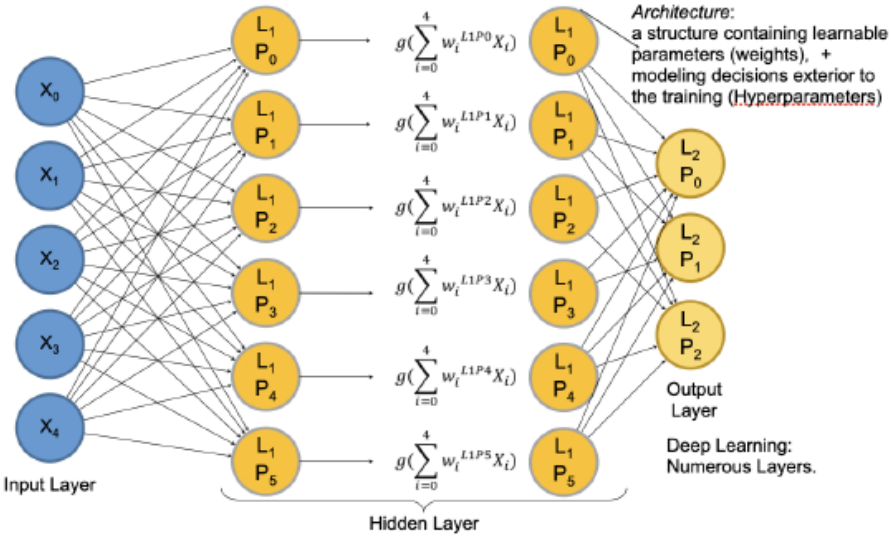
- If classification is incorrect, modify the weight vector w using :

$$w_j(t+1) = w_j(t) - \eta\big(Y_i - \widehat{Y}_i(t)\big)X_{i,j}$$

The perceptron is a linear classifier, therefore it will never get to the state with all the input vectors classified correctly if the training set is not linearly separable, for example, if the positive examples can not be separated from the negative examples by a hyperplane. In this case, no "approximate" solution will be gradually approached under the standard learning algorithm, but instead learning will fail completely.

11.3.2 Single-layer perceptron

We illustrate the Single-layer perceptron or single **hidden** layer Neural Network.



11.3.3 Multilayer perceptron

11.4 Convolutional Neural Networks

11.4.1 Introduction

11.4.2 Convolutional Networks

11.4.3 Dropout and Early stopping

12 Modélisation de séries temporelles (M2)

13 Numerical Finance (M2)

This course aims to connect rigorous probabilistic pricing theory with implementable numerical algorithms, highlighting both the mathematical structure (martingales, change of measure, PDE connections) and the practical engineering constraints (accuracy, stability, computational cost) that determine how derivatives are priced and risk-managed in practice.

The first part establishes the core theoretical tools : stochastic calculus (Itô's formula, stochastic integrals, quadratic variation), self-financing trading strategies, and the role of discounting. We then derive the risk-neutral valuation principle by constructing an equivalent martingale measure via Girsanov's theorem (under appropriate integrability conditions such as Novikov's criterion). This leads to the fundamental pricing representation of contingent claims as discounted conditional expectations under the risk-neutral measure. Completeness and replication in the Black–Scholes model are linked to the martingale representation theorem, providing both theoretical justification and a blueprint for hedging strategies.

13.1 Stochastic Calculus

Let us denotes $(\Omega, \mathcal{F}, \mathbb{P})$ a given probability space endoxed with a completed filtration $(\mathcal{F}_t)_{t\geq 0}$. A family of random variables (X_t) is an $\mathcal{F}_t$-adapted process if

$$\begin{array}{rccc} X & : & \Omega \times \mathbb{R}_+ & \to & \mathbb{R}^d \\ & & (\omega, t) & \mapsto & X_t(\omega) \end{array}$$

such that

- For a fixed $t \in \mathbb{R}_+$, $\omega \mapsto X_t(\omega)$ denotes a random variables.
- For a fixed $\omega \in \Omega$, $t \mapsto X_t(\omega)$ is a path of the process.
- The process (X_t) is a $\mathcal{F}_t$-adapted process if for all $t \geq 0$ X_t is $\mathcal{F}_t$-measurable.

Définition 13.1.1 (Continuous process)**.** If $t \mapsto X_t$ is almost surely continuous, then the process (X_t) is said to be a continuous process.

Définition 13.1.2 (Natural filtration)**.** Given a process (X_t), the natural filtration associated with this process is defined by

$$\mathcal{F}_t = \sigma(X_s | 0 \leq s \leq t)$$

Moreover, if Y is $\sigma(X_s|0 \leq s \leq t)$-measurable then there exists φ such that $Y = \varphi(X_s|0 \leq s \leq t)$.

Définition 13.1.3 (Stopping time)**.** Let $\tau : \Omega \to \overline{\mathbb{R}_+}$ is a given random time. τ is said to be a stopping time with respect to a given filtration if $\{\tau \leq t\}$ is $\mathcal{F}_t$-measurable for all $t \geq 0$.

Définition 13.1.4 (Stopped process)**.** Let (X_t) be a given process, then (X_t^τ) is the stopped process defined by

$$X_t^\tau(\omega) = X_{t\wedge\tau(\omega)}(\omega)$$

Définition 13.1.5 (Brownian Motion)**.** The process (W_t) is a standard brownian motion if

- $W_0 = 0$.

- (W_t) has independant and stationnary increments.
- $\forall t \geq 0$, $W_t \sim \mathcal{N}(0,t)$, ie, $\text{Var}(W_t) = \mathbb{E}(W_t^2) = t$.
- $t \mapsto W_t$ is continuous a.s.

Définition 13.1.6 ($(\mathcal{F}_t)$-Brownian Motion)**.** The process (W_t) is an $(\mathcal{F}_t)$-Brownian Motion for a given filtration $(\mathcal{F}_t)$ if (W_t) is a Brownian Motion satisfying

- (W_t) is $\mathcal{F}_t$-adapted ($\forall t \geq 0,\ W_t \in \mathcal{F}_t$).
- $\forall t, s \geq 0,\ t \geq s,\ W_t - W_s \perp\!\!\!\perp \mathcal{F}_s$.

Définition 13.1.7 (Continuous martingale)**.** A process (M_t) is an $(\mathcal{F}_t)$-martingale for a given filtration $(\mathcal{F}_t)$ if

1. $\forall t \geq 0,\ \mathbb{E}(|M_t|) < \infty$.
2. $\forall t \geq s \geq 0,\ \mathbb{E}(M_t|\mathcal{F}_s) = M_s$.

Same as in the previous chapters, we can define a (super/sub)-martingale.

For a martingale (M_t), $t \mapsto \mathbb{E}(M_t)$ is constant equal to $\mathbb{E}(M_0)$. For a supermartingale, it is decreasing. For a submartingale, it is increasing.

Théorème 13.1.1 (Jensen inequality - Essential)**.** *Let $\varphi \in \mathcal{F}(\mathbb{R},\mathbb{R})$ a convex function. Then for all X random variable on $(\Omega,\mathcal{F},\mathbb{P})$ such that $\mathbb{E}(|\varphi(X)|) < \infty$, we have*

$$\varphi\big(\mathbb{E}(X|\mathcal{G})\big) \leq \mathbb{E}\big(\varphi(X)|\mathcal{G}\big)$$

when $\mathcal{G} \subseteq \mathcal{F}$.

We can then easily show that if (M_t) is a martingale, then (M_t^2) is a submartingale.

Let X be a random variable on $(\Omega,\mathcal{F},\mathbb{P})$ with a given filtration $(\mathcal{F}_t)$. If $X \in L^2(\Omega)$ then

$$\inf_{Z\in\mathcal{F}_s} \mathbb{E}\big(|Z-X|^2\big) = \mathbb{E}\big(|\mathbb{E}(X|\mathcal{F}_s) - X|^2\big)$$

$L^2(\Omega)$ is a Hilbert space (I advice to read the lesson on normed vectorial spaces and differential calculus by Karine Beauchard for more explanations) endowed with the inner product

$$\langle X,Y\rangle_{L^2(\Omega)} = \mathbb{E}(XY)$$

and the implied norm

$$||X||_{L^2(\Omega)} = \sqrt{\langle X,Y\rangle_{L^2(\Omega)}}$$

And we know that, $X \perp\!\!\!\perp Y$ if $\langle X,Y\rangle_{L^2(\Omega)} = \mathbb{E}(XY) = 0$. By considering the $\mathcal{F}_s$ plane and the fact that $\mathbb{E}(X|\mathcal{F}_s)$ is the orthogonal projection of X on $\mathcal{F}_s$, we get that $\mathbb{E}\big((X - \mathbb{E}(X|\mathcal{F}_s))Z\big) = 0$ for Z $\mathcal{F}_s$-measurable.

Proposition 13.1.1. *With a Brownian motion (W_t) we can build 3 types of martingales*

1. (W_t).
2. $(W_t^2 - t)$.
3. $\Big(exp\big(\lambda W_t - \frac{\lambda^2}{2}t\big)\Big)$, *for* $\lambda \in \mathbb{R}$.

Théorème 13.1.2 (Doob inequality - Useful tool)**.** *If* (M_t) *is a given continuous* L^2*-martingale then*

$$\mathbb{E}\big(\sup_{0\leq t\leq T}|M_t|^2\big) \leq 4\mathbb{E}(|M_T|^2)$$

Now that we reformulated the generalities and the main definitions, we can reintroduce the most important tool in stochastic calculus, or at least in the financial world : the Itô integral.

Let us define for a given filtration $(\mathcal{F}_t)$, the set

$$\mathcal{H}_T = \left\{(H_t), (\mathcal{F}_t)\text{-adapted processes} \;\middle|\; \mathbb{E}\Big(\int_0^T |H_s|^2 ds\Big) < \infty\right\}$$

Définition 13.1.8 (Itô integral)**.** Let $(H_t) \in \mathcal{H}_T$ a given process, then the process denoted $\Big(\int_0^t H_s dW_s\Big)$ is defined as a martingale satisfying the isometry property

$$\mathbb{E}\Big(\Big(\int_0^T H_s dW_s\Big)^2\Big) = \mathbb{E}\Big(\int_0^T H_s^2 ds\Big)$$

One have to keep in mind that

$$\int_0^T H_s dW_s = \lim_{n\infty}\sum_{k=0}^{n-1} H_{t_k}(W_{t_{k+1}} - W_{t_k})$$

where $0 = t_0 < t_1 < ... < t_n = T$.

Thanks to that formulation, we can derive the usual following properties :

- $\displaystyle\int_0^T (\alpha H_s^1 + \beta H_s^2) dW_s = \alpha \int_0^T H_s^1 dW_s + \beta \int_0^T H_s^2 dW_s$
- $\forall 0 \leq t \leq T,\ {}_0^T H_s\, dW_s = \displaystyle\int_0^t H_s\, dW_s + \int_t^T H_s\, dW_s$
- $\displaystyle\int_0^0 H_s dW_s = 0.$

Now, let us define the set

$$\tilde{\mathcal{H}}_T = \left\{(H_t), (\mathcal{F}_t)\text{-adapted processes} \;\middle|\; \int_0^T H_s^2 ds < \infty \; a.s.\right\}$$

Définition 13.1.9 (Local martingale)**.** A process (M_t) is a local martingale if there exists an increasing sequence of stopping time $\tau_n \to \infty$ such that the stopped process $(M_{t\wedge\tau_n})$ is an $(\mathcal{F}_t)$-martingale.

Proposition 13.1.2. *If (H_t) is a process belonging to $\tilde{\mathcal{H}}_T$ then $\left(\int_0^t H_s dW_s\right)$ is well-deined as a local martingale.*

When $(H_t) \in \tilde{\mathcal{H}}_T$, we loose the isometry property except if we consider the stopped stochastic integral.

Définition 13.1.10 (Covariation)**.** Guven two processes (X_t) and (Y_t), the covariation process of X and Y denoted by $\left([X,Y]_t\right)$ is defined, if the below limit exists, as :

$$[X,Y]_t = \lim_{n\infty} \sum_{i=1}^{n} (X_{t_i\wedge t} - X_{t_{i-1}\wedge t})(Y_{t_i\wedge t} - Y_{t_{i-1}\wedge t})$$

13.2 The Black-Scholes model

13.3 Pricing and Hedging portfolio under Black-Scholes

13.4 Stochastic and Partial differential equations

13.5 Local and Stochastic volatility models

13.6 Pricing under Stochastic volatility models

13.7 Discretization schemes for SDEs

13.8 Lévy process and applications

14 Advanced Process Approximation (M2)

15 Particule system and McKean Vlasov SDE and application to Machine Learning (M2)