Probabilités et statistiques

Paul Viallard
(paul.viallard@inria.fr)

2025

Table des matières

Avant-propos

Bibliographie

Ce cours est basé sur celui de François Schwarzentruber, qui est lui-même basé sur différentes références. Des notations inspirées de mes papiers et de ceux de Pascal Germain, François Laviolette, Mario Marchand, et d'Emilie Morvant.

Références pour la théorie des probabilités

Le cours de François, sur lequel ce cours est basé, s'inspire de deux références pour les probabilités : (Garet et al., 2019) et (Barbe et al., 2021). Cependant, ce cours ajoute des éléments sur la théorie de la mesure qui ne proviennent pas de ces livres. De plus, voici deux livres qui sont intéressants (en tout cas, qui m'intéressent particulièrement...) : (Whittle, 2012 ; Shafer et al., 2019).

Références pour l'apprentissage statistique

Pour une compréhension globale de l'apprentissage statistique, plusieurs références peuvent être intéressantes. Parmi les ouvrages présentant les méthodes d'un point de vue pratique, (Bishop et al., 2006) et (Hastie et al., 2009) sont des références incontournables ; il y a également des références plus récentes comme (Murphy, 2022). D'un point de vue théorique, il y a le livre de (Vapnik, 2013), qui jette les bases théoriques des principes fondamentaux du domaine. Enfin, des introductions plus récentes de ces bases théoriques se trouvent dans (Mohri et al., 2012) et (Bach, 2024).

Partie I

Théorie des probabilités

Chapitre 1

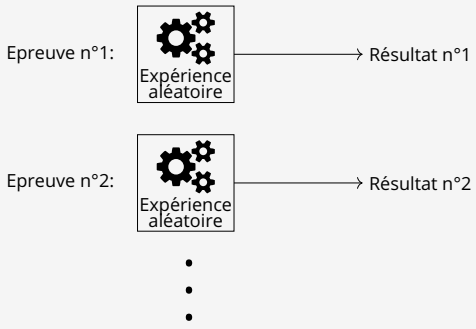
Mesures et espaces probabilisés

1.1 Introduction

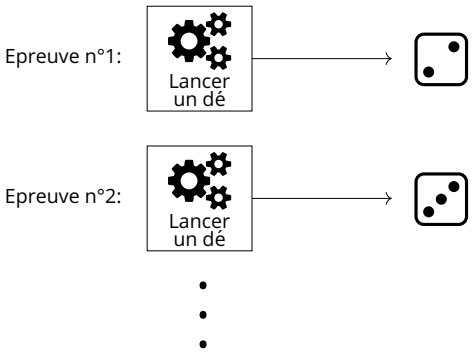
Expérience aléatoire

La théorie des probabilités sert à étudier une expérience aléatoire.

Définition 1.1. Une *expérience aléatoire* est un processus dont le résultat ne peut pas être déterminé à l'avance, même si les conditions initiales sont connues. Autrement dit, le résultat est incertain et dépend du hasard.



Exemple. *Le lancer de dé est une expérience aléatoire :*



Caractérisation d'une expérience aléatoire

Afin de caractériser une expérience aléatoire, nous avons besoin de définir

- de tous les résultats pouvant être obtenus,
- les probabilités associées à chaque résultat pour quantifier la "chance" d'observer un résultat donné.

Exemple. *Pour un lancer de dé, nous avons*

- *l'ensemble des résultats défini par* $\{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$,

- *les probabilités définies par* $\mathbb{P}_{\text{dé}\sim\text{lancer}}\left[\text{dé} = \boxdot\right] = \frac{1}{6}$, $\mathbb{P}_{\text{dé}\sim\text{lancer}}\left[\text{dé} = \boxdot\right] = \frac{1}{6}$, $\mathbb{P}_{\text{dé}\sim\text{lancer}}\left[\text{dé} = \boxdot\right] = \frac{1}{6}$, $\mathbb{P}_{\text{dé}\sim\text{lancer}}\left[\text{dé} = \boxdot\right] = \frac{1}{6}$, $\mathbb{P}_{\text{dé}\sim\text{lancer}}\left[\text{dé} = \boxdot\right] = \frac{1}{6}$, *et* $\mathbb{P}_{\text{dé}\sim\text{lancer}}\left[\text{dé} = \boxdot\right] = \frac{1}{6}$.

Objectif du chapitre

L'objectif de ce chapitre est de définir formellement la notion d'expérience aléatoire. Pour ce faire, nous introduirons les espaces probabilisés (qui formaliseront la notion d'expérience aléatoire) en nous appuyant sur des éléments de théorie de la mesure.

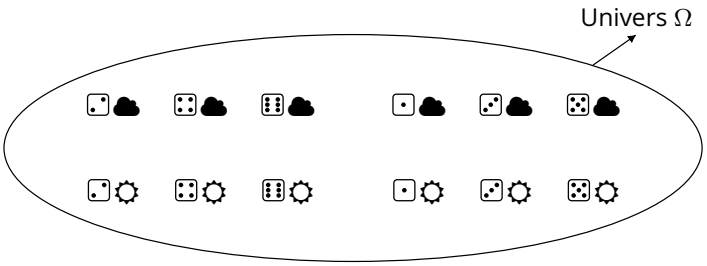
1.2 Elements de la théorie des probabilités

Univers

Afin de caractériser l'ensemble des résultats possibles d'une expérience aléatoire, nous devons définir l'univers.

Définition 1.2. Un *univers* Ω est l'ensemble (non vide) de tous les résultats pouvant être obtenus au cours d'une expérience aléatoire.

Exemple. *Voici un exemple d'univers* Ω *avec 12 résultats possibles :*

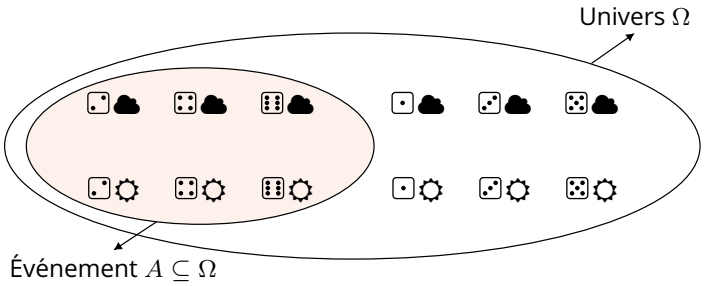


Événement

Parfois, nous pouvons nous intéresser qu'à un sous-ensemble de Ω. Pour cela, nous devons définir un événement.

Définition 1.3. Un *événement* A est un sous-ensemble de Ω désignant un sous-ensemble des résultats possibles d'une expérience aléatoire.

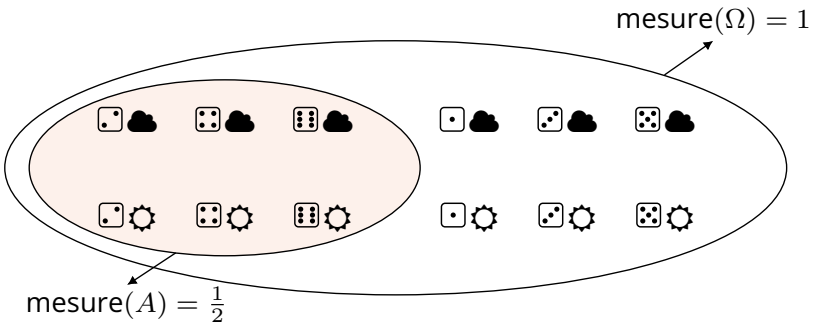
Exemple. *Voici l'événement* $A :=$ *"le dé affiche un nombre pair"*



Pour définir des probabilités...

Pour définir des probabilités, nous allons mesurer la "taille" des événements $A \subseteq \Omega$ à l'aide d'un objet mathématique appelé "mesure". Ainsi, nous verrons que définir une mesure revient à définir des probabilités.

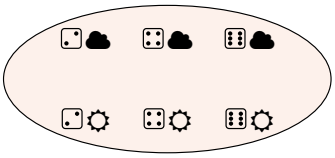
Exemple. *Voici un exemple de mesure sur l'univers* Ω *et l'événement* $A \subseteq \Omega$ *précédent.*



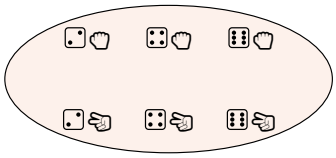
Pour définir une mesure, nous devons également identifier les différents ensembles autorisés à être mesurés.

Exemple. *Nous voulons autoriser que seuls les événements $A \subseteq \Omega$ puissent être mesurés.*

On peut mesurer l'événement $A \subseteq \Omega$



On ne peut pas mesurer l'ensemble $A \nsubseteq \Omega$



Introduction (rapide) à la théorie de la mesure

Afin de bien définir une mesure, nous allons voir différentes notions issues de la théorie de la mesure.

Tribu (ou σ-algèbre)

Pour définir l'ensemble des ensembles que l'on peut mesurer, nous allons définir une tribu.

Définition 1.4. Une *tribu* Σ_X (ou σ-algèbre) est un ensemble de sous-ensembles de X avec :

1. $\emptyset \in \Sigma_X$;
2. $X \in \Sigma_X$;
3. Σ_X est stable par complémentaire : si $A \in \Sigma_X$, alors $\overline{A} \in \Sigma_X$;
4. Σ_X est stable par union dénombrable : si $\forall n \in \mathbb{N}, A_n \in \Sigma_X$, alors $\cup_{n\in\mathbb{N}} A_n \in \Sigma_X$;
5. Σ_X est stable par intersection dénombrable : si $\forall n \in \mathbb{N}, A_n \in \Sigma_X$, alors $\cap_{n\in\mathbb{N}} A_n \in \Sigma_X$.

Définition 1.5. Si un ensemble X est muni d'une tribu Σ_X, alors (X, Σ_X) est appelé *espace mesurable*.

Mesure

Nous sommes maintenant prêts à définir une mesure : c'est une fonction ν (positive), qui mesure la "taille" d'un ensemble A appartenant à la tribu Σ_X.

Définition 1.6. Une *mesure* est une application ν définie sur la tribu Σ_X et à valeur dans $[0, +\infty]$ qui respecte les deux propriétés suivantes :

- $\nu(\emptyset) = 0$
- Pour toute suite de sous-ensembles $(A_i)_{i\in\mathbb{N}}$ de Σ_X deux à deux disjoints, on a

$$\nu\left(\bigsqcup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} \nu(A_i).$$

Définition 1.7. Si un ensemble X est muni d'une tribu Σ_X, alors (X, Σ_X, ν) est appelé *espace mesuré*.

La proposition qui suit présente des propriétés naturelles sur un espace mesuré.

Proposition 1.8. Si (X, Σ_X, ν) est un espace mesuré où $\nu : \Sigma_X \to \mathbb{R}^+$, nous avons les propriétés suivantes :

1. Pour tout $A, B \in \Sigma_X$, $A \cap B = \emptyset$ implique $\nu(A \sqcup B) = \nu(A) + \nu(B)$;
2. Pour tout $A, B \in \Sigma_X$, $\nu(A \cup B) = \nu(A) + \nu(B) - \nu(A \cap B)$;
3. Pour tout $A, B \in \Sigma_X$, $\nu(A \cup B) \leq \nu(A) + \nu(B)$;
4. Pour tout $A, B \in \Sigma_X$, $\nu(A \cap B) \leq \min(\nu(A), \nu(B))$;
5. Pour tout $A, B \in \Sigma_X$, $\nu(A \cup B) \geq \max(\nu(A), \nu(B))$.

1.3 Théorie de la mesure appliquée à la théorie des probabilités

Tribu sur un univers

Pour définir des probabilités, un univers Ω doit être muni d'une tribu.

Définition 1.9. Si un univers Ω est muni d'une tribu Σ_Ω, alors (Ω, Σ_Ω) est appelé *espace probabilisable*.

Définition 1.10. Un événement est un sous-ensemble A qui est contenu dans Σ_Ω.

Mesure de probabilité

Ensuite, une mesure de probabilité est une fonction μ, qui, à tout événement A, associe un nombre $\mu(A)$ entre 0 et 1, qui mesure de combien l'événement est probable.

Définition 1.11. Une *mesure de probabilité* est une application de $\mu : \Sigma_\Omega \to [0, 1]$ vérifiant :

- $\mu(\Omega) = 1$;
- $\mu(\emptyset) = 0$;
- Pour toute suite de sous-ensembles $(A_i)_{i \in \mathbb{N}}$ de Σ_Ω deux à deux disjoints, on a

$$\mu\left(\bigsqcup_{i \in \mathbb{N}} A_i\right) = \sum_{i \in \mathbb{N}} \mu(A_i).$$

Plus précisément, $\mu(\emptyset) = 0$ car nous mesurons un événement sans aucun résultat possible. De même, comme Ω est l'ensemble de tous les résultats possibles, nous avons $\mu(\Omega) = 1$. On peut aussi imaginer que Ω est une surface d'aire 1. Alors $\mu(A)$ est l'aire de A.
Nous verrons dans le prochain chapitre que la mesure de probabilité nous permet de définir directement des probabilités. Autrement dit, le nombre $\mu(A)$ est la probabilité que A soit vraie.

Proposition 1.12. Soit $(\Omega, \Sigma_\Omega, \mu)$ un espace probabilisé. Nous avons

$$\mu(A) = \mathop{\mathbb{P}}_{\omega \sim \mu} [\omega \in A] .$$

Espace probabilisé

Un espace probabilisé est constitué d'un univers Ω, d'une tribu définie sur cet univers, et d'une mesure de probabilité.

Définition 1.13. Un *espace probabilisé* est $(\Omega, \Sigma_\Omega, \mu)$ est la donnée :

- d'un espace Ω ;
- d'une tribu Σ_Ω sur Ω ;
- d'une mesure de probabilité $\mu : \Sigma_\Omega \to [0, 1]$.

En plus des propriétés de Proposition 1.8, nous avons la propriété suivante pour un espace probabilisé.

Proposition 1.14. Si $(\Omega, \Sigma_\Omega, \mu)$ est un espace probabilisé, nous avons :

1. Pour tout $A \in \Sigma_\Omega$, $\mu(\overline{A}) = 1 - \mu(A)$.

Voici deux exemples simples d'espaces probabilisés où Ω est un ensemble fini.

Exemple (Lancer d'un dé)**.**

- $\Omega = \{1, 2, 3, 4, 5, 6\}$;
- $\Sigma_\Omega = 2^{\{1,2,3,4,5,6\}}$;
- *Pour tout* $A \subseteq \Omega$*, nous avons* $\mu(A) = \frac{|A|}{6}$.

Exemple (Météo)**.**

- $\Omega = \{$☼, 🌥$\}$;
- $\Sigma_\Omega = 2^{\{$☼, 🌥$\}}$;
- $\mu(\emptyset) = 0$, $\mu(\Omega) = 1$, $\mu(\{$☼$\}) = 0.2$, $\mu(\{$🌥$\}) = 0.8$.

Par contre, lorsque l'on veut définir une mesure de probabilité sur un intervalle, tout se complique ... En effet, nous pouvons prouver le théorème suivant.

Théorème 1.15. Soit $\Omega = [0, 1]$ et $\Sigma_\Omega = 2^\Omega$, alors il n'existe pas d'espace probabilisé $(\Omega, \Sigma_\Omega, \mu)$ tel que $\mu([a, b]) = b - a$ avec $0 \leq a \leq b \leq 1$.

1.4 Tribu borélienne et mesure de Lebesgue

Autrement dit, nous ne pouvons pas définir de mesure de probabilité μ naturelle sur un intervalle (où sa mesure correspond à la longueur de l'intervalle) si la tribu est l'ensemble des parties de Ω.

Tribu engendrée et tribu borélienne

Pour réussir à définir une mesure μ sur $[0, 1]$, il faut utiliser une autre tribu : la tribu borélienne.
Pour la construire, il faut supposer que

- $[a, b]$ avec $a, b \in [0, 1]$ est dans *Borel*$([a, b])$.

Puis, *Borel*$([a, b])$ est stable par complémentaire, union, et intersection dénombrables. Nous avons, par exemple

- $[0, 1] \setminus [a, b]$ est dans *Borel*$([0, 1])$;
- $\bigcup_{i\in\mathbb{N}}[a_i, b_i]$ est dans *Borel*$([0, 1])$ avec $0 \leq a_i \leq b_i \leq 1$;
- $\bigcap_{i\in\mathbb{N}}[a_i, b_i]$ est dans *Borel*$([0, 1])$ avec $0 \leq a_i \leq b_i \leq 1$;
- $\{a\} = [a, a]$ est dans *Borel*$([0, 1])$;
- Tout ensemble dénombrable $X = \bigcup_{x\in X}\{x\}$ est dans *Borel*$([0, 1])$.

Les deux définitions suivantes formalisent exactement cette construction.

Définition 1.16. Soit A un ensemble de sous-ensembles de Ω. La *tribu engendrée* par A est la plus petite tribu contenant A.

Définition 1.17. La *tribu borélienne Borel*$([0, 1])$ est la tribu engendrée par l'ensemble $\{[a, b] \mid a, b \in [0, 1]\}$.

Tribu borélienne sur un ensemble $F \subseteq \mathbb{R}^d$

Nous allons voir que nous pouvons étendre la notions de tribu borélienne à un ensemble $F \subseteq \mathbb{R}^d$, avec $d \in \mathbb{N}^*$. Ceci nous servira plus tard dans le cours.

Définition 1.18. La *tribu borélienne Borel*(F) où $F \subseteq \mathbb{R}^d$, avec $d \in \mathbb{N}^*$, est la tribu engendrée par l'ensemble

$$\left\{ \prod_{i=1}^{d} [a_i, b_i] \;\middle|\; \prod_{i=1}^{d} [a_i, b_i] \subseteq F \right\}$$

Mesure de Lebesgue

Nous pouvons définir la mesure de Lebesgue, qui est une mesure sur un produit cartésien d'intervalles provenant de *Borel*(F).

Définition 1.19. Soit I_i un intervalle de la forme $[a_i, b_i]$, $]a_i, b_i]$, $[a_i, b_i[$, ou $]a_i, b_i[$. La *mesure de Lebesgue* ν d'un produit cartésien d'intervalles est donnée par

$$\nu(I_1 \times \cdots \times I_d) := \prod_{i=1}^{d} (b_i - a_i).$$

Espace probabilisé avec une tribu borélienne et une mesure de Lebesgue

Grâce à la tribu borélienne *Borel*$([0,1])$ et sa mesure de Lebesgue associée, nous pouvons définir un espace probabilisé sur l'interval $[0,1]$.

Exemple (Loi uniforme sur un intervalle)**.**

- $\Omega = [0,1]$ *;*
- $\Sigma_\Omega =$ *Borel*$([0,1])$ *;*
- $\mu([a,b]) = \mu(]a,b]) = \mu([a,b[) = \mu(]a,b[) = b - a$ *avec* $0 \leq a \leq b \leq 1$.

Chapitre 2

De l'intégration aux probabilités

Introduction

Dans le chapitre précédent, nous avons vu avec Proposition 1.12 l'égalité suivante.

Soit $(\Omega, \Sigma_\Omega, \mu)$ un espace probabilisé, alors nous avons

$$\mu(A) = \mathop{\mathbb{P}}_{\omega\sim\mu}[\omega \in A].$$

Dans ce chapitre, nous allons démontrer cette proposition.
De plus, nous allons définir des notions essentielles pour caractériser une expérience aléatoire. En effet, étant donné un espace probabilisé $(\Omega, \Sigma_\Omega, \mu)$, nous allons définir une probabilité $\mathbb{P}_{\omega\sim\mu}[\cdot]$, une espérance $\mathbb{E}_{\omega\sim\mu}[\cdot]$, et voir certaines de leurs propriétés.
Afin de formaliser ces notions fondamentales en théorie des probabilités, il est indispensable de définir l'intégrale de Lebesgue. Ainsi, nous verrons : *(i)* les fonctions simples, qui servent d'éléments de base pour la définition de l'intégrale, et *(ii)* la notion de fonction mesurable, c'est-à-dire une fonction pour laquelle cette intégrale peut être correctement définie.

2.1 Fonction simple

Fonction simple

Pour définir l'intégrale de Lebesgue, nous avons besoin de définir un type de fonction particulier.

Définition 2.1. Soit (X, Σ_X) un espace mesurable. Soient $a_1, \dots, a_n \in \mathbb{R}$ et $A_1, \dots, A_n \in \Sigma_X$ des ensembles mesurables deux à deux disjoints et $\cup_{i\in\{1,\dots,n\}} A_n = A \subseteq X$. Une *fonction simple* $f : A \to \mathbb{R}$ est définie par

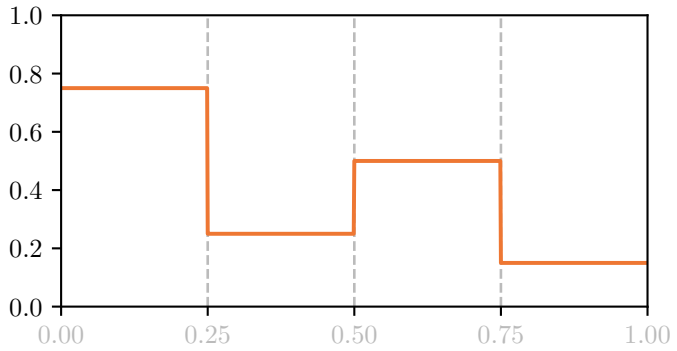
$$\forall x \in A, \quad f(x) = \sum_{k=1}^{n} a_k \mathbb{1}[x \in A_k],$$

où $\mathbb{1}[a] = 1$ si la proposition a est vraie et 0 sinon.

En d'autres termes, une fonction simple prend un nombre fini de valeurs réelles. De plus, pour tout élément $x \in A_k$, la fonction simple donne la valeur associée à l'ensemble A_k, c'est-à-dire que nous avons $f(x) = a_k$.
Voici un exemple de fonction simple, ainsi que sa représentation graphique.

Exemple.

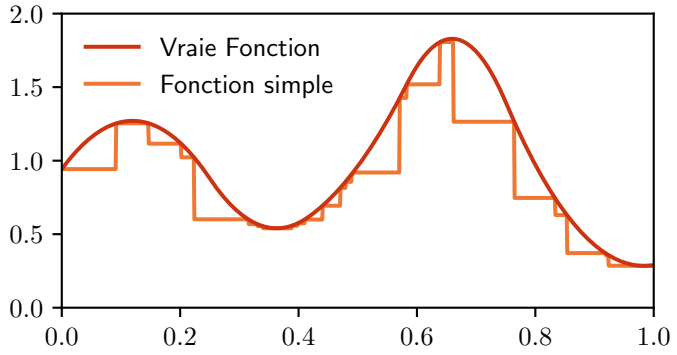
$$f(x) = 0.75 \cdot \mathbb{1}\Big[x \in [0.0, 0.25[\Big] + 0.25 \cdot \mathbb{1}\Big[x \in [0.25, 0.5[\Big] + 0.5 \cdot \mathbb{1}\Big[x \in [0.5, 0.75[\Big] + 0.15 \cdot \mathbb{1}\Big[x \in [0.75, 1.0]\Big]$$



Approximation par fonctions simples

En pratique, une fonction simple permet d'approcher une fonction (plus compliquée).

Exemple.



Cette approximation est justifiée par un théorème que nous allons voir plus tard dans ce chapitre. Ainsi, le théorème suivant n'est que sa version informelle.

Théorème (informel !). Pour un certain type de fonction $f : X \to \mathbb{R}$, il existe une séquence de fonctions simples $\{f_n\}_{n\in\mathbb{N}}$ telle que

$$\text{pour tout } x \in X, \quad \text{nous avons} \quad \lim_{n\to\infty} f_n(x) = f(x).$$

En réalité, il permet d'approcher des fonctions mesurables (que nous allons voir ensuite).

2.2 Fonction mesurable

Fonction mesurable

L'intégrale de Lebesgue permet d'intégrer un type de fonction, appelé fonction mesurable, que nous allons voir dans la définition suivante.

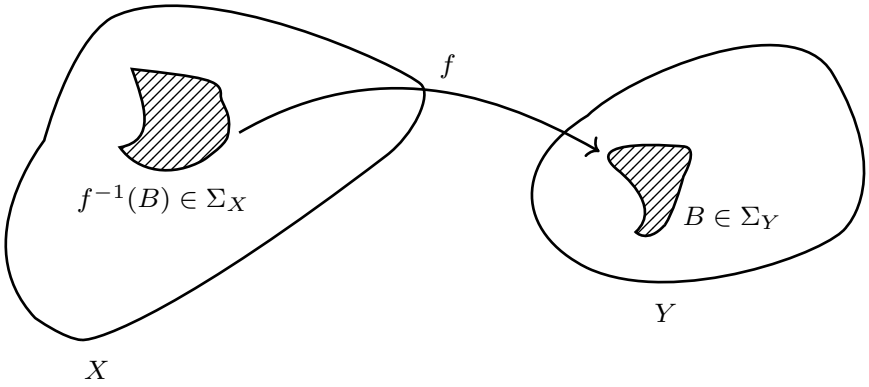
Définition 2.2. Soient deux espaces (X, Σ_X) et (Y, Σ_Y) mesurables. Une fonction $f : X \to Y$ est dite *mesurable* si nous avons

$$\text{pour tout } B \in \Sigma_Y, \quad \text{nous avons} \quad f^{-1}(B) \in \Sigma_X,$$

$$\text{où} \quad f^{-1}(B) = \{x \in X \mid f(x) \in B\}.$$

Voici une illustration pour clarifier cette définition.

Exemple.



Cette définition garantit que la fonction f "transporte" les ensembles mesurables de Y (c'est-à-dire $B \in \Sigma_Y$) vers des ensembles mesurables de X (c'est-à-dire $f^{-1}(B) \in \Sigma_X$). Autrement dit, elle assure que f préserve la structure mesurable des espaces X et Y, ce qui permet d'étudier ses propriétés à travers l'intégration.

Propriétés des fonctions mesurables

Si l'espace Y est restreint à l'ensemble des réels $\mathbb{R}$, nous pouvons trouver une caractérisation (plus naturelle) des fonctions mesurables.

Théorème 2.3. Soient (X, Σ_X) un espace mesurable et une fonction $f : X \to \mathbb{R}$. Les propositions suivantes sont équivalentes

1. f est une fonction mesurable,
2. pour tout $\alpha \in \mathbb{R}$, nous avons $\{x \in X \mid f(x) \leq \alpha\} \in \Sigma_X$,
3. pour tout $\alpha \in \mathbb{R}$, nous avons $\{x \in X \mid f(x) < \alpha\} \in \Sigma_X$,
4. pour tout $\alpha \in \mathbb{R}$, nous avons $\{x \in X \mid f(x) \geq \alpha\} \in \Sigma_X$,
5. pour tout $\alpha \in \mathbb{R}$, nous avons $\{x \in X \mid f(x) > \alpha\} \in \Sigma_X$.

En effet, pour une fonction $f : X \to \mathbb{R}$ qui est mesurable, il suffit de vérifier que les ensembles où la fonction est supérieure ou inférieure à un certain seuil α sont mesurable, c'est-à-dire qu'ils appartiennent à la tribu Σ_X. Autrement dit, cette propriété reflète le fait que f "transporte" les ensembles mesurables de $\mathbb{R}$ (comme les intervalles $]-\infty, \alpha]$ ou $[\alpha, +\infty[$) vers des ensembles mesurables de X.
Grâce à cette propriété, nous pouvons prouver que les fonctions simples sont des fonctions mesurables.

Théorème 2.4. Soient un espace mesurable (X, Σ_X) et $f : X \to \mathbb{R}$ une fonction simple, alors f est une fonction mesurable.

De plus, une caractérisation importante des fonctions mesurables est qu'elles peuvent être approchées par des fonctions simples. En effet, nous avons le théorème suivant.

Théorème 2.5. Soient un espace mesurable (X, Σ_X) et $f : X \to \mathbb{R}$ une fonction mesurable, alors il existe une séquence de fonctions simples $\{f_n\}_{n\in\mathbb{N}}$ telle que

$$\text{pour tout } x \in X, \quad \text{nous avons} \quad \lim_{n\to\infty} f_n(x) = f(x).$$

Autrement dit, nous pouvons approcher n'importe quelle fonction mesurable f à l'aide d'une séquence $\{f_n\}_{n\in\mathbb{N}}$ de fonctions simples.

2.3 Intégrale de Lebesgue

Maintenant, nous pouvons définir l'intégrale de Lebesgue pour les fonctions mesurables. Nous allons tout d'abord nous concentrer sur les fonctions simples (qui sont des fonctions mesurables).

Définition 2.6. Soient un espace mesuré (X, Σ_X, ν) et $A \in \Sigma_X$, soit une fonction simple $f : A \to \mathbb{R}$ où

$$\text{pour tout} \quad x \in A, \quad \text{nous avons} \quad f(x) = \sum_{k=1}^{n} a_k \mathbb{1}[x \in A_k].$$

L'*intégrale de Lebesgue pour une fonction simple* f est définie par

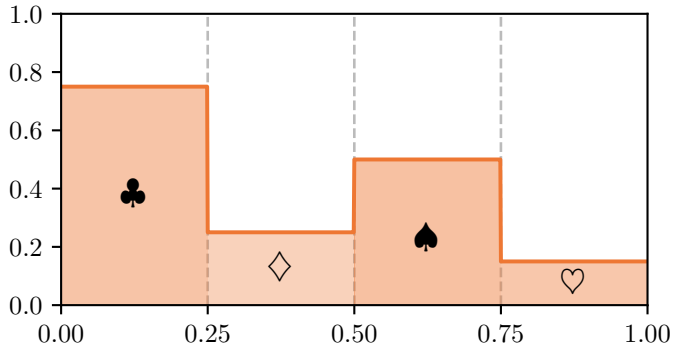
$$\int_A f(x) d\nu(x) := \sum_{k=1}^{n} a_k \nu(A_k).$$

De façon non formelle, lorsque nous avons une fonction simple $f : X \to \mathbb{R}$, l'intégrale de Lebesgue revient à sommer les aires des rectangles, avec pour "longueur" a_k et "largeur" $\nu(A_k)$.
Voici un exemple d'intégration avec la fonction simple précédente.

Exemple.

$$\int_A f(x) d\nu(x) = \underbrace{0.75 \cdot \nu([0.0, 0.25[)}_{\clubsuit} + \underbrace{0.25 \cdot \nu([0.25, 0.5[)}_{\diamondsuit} + \underbrace{0.5 \cdot \nu([0.5, 0.75[)}_{\spadesuit} + \underbrace{0.15 \cdot \nu([0.75, 1.0])}_{\heartsuit}$$

(où ν est la mesure de Lebesgue)



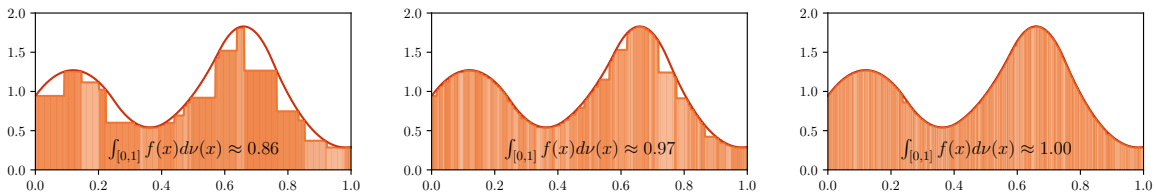
Maintenant que nous avons une intégrale pour les fonctions simples, nous pouvons les définir pour des fonctions plus générales qui sont les fonctions mesurables positives.

Définition 2.7. Soient un espace mesuré (X, Σ_X, ν) et $A \in \Sigma_X$, l'*intégrale de Lebesgue pour une fonction* $f : A \to \mathbb{R}^+$ *mesurable (positive)* est définie par

$$\int_A f(x) d\nu(x) := \sup \left\{ \int_A s(x) d\nu(x) \;\middle|\; s \text{ fonction simple } f : A \to \mathbb{R} \text{ avec } 0 \leq s(x) \leq f(x) \text{ pour tout } x \in A \right\}.$$

En d'autres termes, l'intégrale de Lebesgue d'une fonction mesurable positive f est obtenue en approximant f par des fonctions simples s telles que $0 \leq s(x) \leq f(x)$ pour tout $x \in X$. L'intégrale de f est alors définie comme la borne supérieure des intégrales de ces fonctions simples s.
Voici une illustration de l'intégration d'une fonction mesurable utilisée dans un exemple précédent.

Exemple.



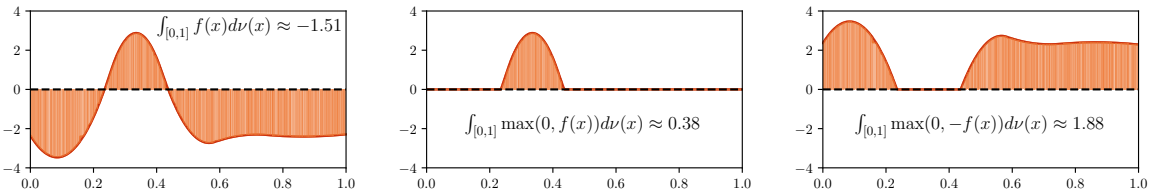
Malheureusement, l'intégrale est définie pour des fonctions mesurables positives. Ainsi, nous pouvons étendre la définition de l'intégrale de Lebesgue aux fonctions mesurables.

Définition 2.8. Soient un espace mesuré (X, Σ_X, ν) et $A \in \Sigma_X$, l'*intégrale de Lebesgue pour une fonction* $f : X \to \mathbb{R}$ *mesurable* est définie par

$$\int_A f(x)d\nu(x) := \int_A \max(0, f(x))d\nu(x) - \int_A \max(0, -f(x))d\nu(x).$$

L'idée est d'utiliser deux intégrales de Lebesgue sur deux fonctions mesurables positives $x \mapsto \max(0, f(x))$ et $x \mapsto \max(0, -f(x))$ et de soustraire la valeur de l'intégrale associée à $x \mapsto \max(0, -f(x))$.
Voici un exemple d'intégration pour une fonction mesurable.

Exemple.



Après avoir défini l'intégrale de Lebesgue pour une fonction mesurable, il est naturel de préciser la condition sous laquelle l'intégration se déroule correctement, c'est-à-dire lorsque l'intégrale prend une valeur finie.

Définition 2.9. Soit un espace mesuré (X, Σ_X, ν) et une fonction $f : X \to \mathbb{R}$ mesurable. Si

$$\int_X |f(x)|d\nu(x) < +\infty$$

alors la fonction f est dite *intégrable* (au sens de Lebesgue).

Nous pouvons également dériver les propriétés suivantes pour l'intégrale de Lebesgue.

Théorème 2.10. Soient un espace mesuré (X, Σ_X, ν) et $f : X \to \mathbb{R}$ et $g : X \to \mathbb{R}$ des fonctions intégrables,

1. pour tout $c \in \mathbb{R}$,
$$\int_X (c \cdot f(x))\, d\nu(x) = c \cdot \int_X f(x)d\nu(x),$$
2. nous avons
$$\int_X (f(x) + g(x))\, d\nu(x) = \int_X f(x)d\nu(x) + \int_X g(x)d\nu(x),$$
3. pour tout $A \in \Sigma_X$,
$$\int_X \mathbb{1}[x \in A]f(x)d\nu(x) = \int_A f(x)d\nu(x).$$

2.4 Espérance et probabilité

Espérance

Grâce à l'intégrale de Lebesgue, nous pouvons maintenant définir l'espérance.

Définition 2.11. Soit un espace probabilisé $(\Omega, \Sigma_\Omega, \mu)$, l'espérance de $\omega \mapsto f(\omega)$ est définie par

$$\mathop{\mathbb{E}}_{\omega\sim\mu}\Big[f(\omega)\Big] := \int_\Omega f(\omega)d\mu(\omega).$$

Grâce à la définition de l'espérance, nous pouvons dériver les propriétés suivantes qui sont directement issues de l'intégrale de Lebesgue.

Corollaire 2.12. Soient un espace probabilisé $(\Omega, \Sigma_\Omega, \mu)$ et $f, g : \Omega \to \mathbb{R}$ des fonctions intégrables,

1. pour tout $c \in \mathbb{R}$, nous avons
$$\mathop{\mathbb{E}}_{\omega\sim\mu}[c \cdot f(\omega)] = c \cdot \mathop{\mathbb{E}}_{\omega\sim\mu}[f(\omega)]$$
2. pour tout $c \in \mathbb{R}$, nous avons
$$\mathop{\mathbb{E}}_{\omega\sim\mu}[f(\omega) + g(\omega)] = \mathop{\mathbb{E}}_{\omega\sim\mu}[f(\omega)] + \mathop{\mathbb{E}}_{\omega\sim\mu}[g(\omega)]$$

Probabilité

De plus, grâce à l'espérance, nous pouvons également définir formellement une probabilité.

Définition 2.13. Soit un espace probabilisé $(\Omega, \Sigma_\Omega, \mu)$, la probabilité que $\omega \in A$ où $\omega \sim \mu$ est définie par
$$\mathop{\mathbb{P}}_{\omega\sim\mu}[\omega \in A] := \mathop{\mathbb{E}}_{\omega\sim\mu} \mathbb{1}[\omega \in A]$$

Enfin, nous pouvons relier une probabilité à sa mesure (grâce à la définition d'une probabilité et de l'espérance). Autrement dit, nous sommes maintenant prêts à prouver Proposition 1.12.

Soit $(\Omega, \Sigma_\Omega, \mu)$ un espace probabilisé. Nous avons
$$\mu(A) = \mathop{\mathbb{P}}_{\omega\sim\mu}[\omega \in A].$$

Démonstration. Grâce aux Définitions 2.13 et 2.11, nous avons
$$\begin{aligned}\mathop{\mathbb{P}}_{\omega\sim\mu}[\omega \in A] &= \mathop{\mathbb{E}}_{\omega\sim\mu} \mathbb{1}[\omega \in A] \\ &= \int_\Omega \mathbb{1}[\omega \in A]\, d\mu(\omega).\end{aligned}$$

De plus grâce à Théorème 2.10 (propriété 3) et Définition 2.6, nous avons
$$\begin{aligned}\mathop{\mathbb{P}}_{\omega\sim\mu}[\omega \in A] &= \int_\Omega \mathbb{1}[\omega \in A]\, d\mu(\omega) \\ &= \int_A 1 d\mu(\omega) \\ &= \mu(A).\end{aligned}$$

□

Chapitre 3

Propriétés de l'espérance

Introduction

Nous avons vu dans le chapitre précédent la définition de l'espérance et d'une probabilité. Nous allons donc voir différentes propriétés. Plus précisément, nous allons aller un peu plus loin que les propriétés naturelles du Théorème 2.10. En effet, nous allons voir des propriétés que l'on rencontre souvent en probabilités et statistiques. Nous allons voir

- *La dérivée de Radon-Nikodym*
- *Les variables aléatoires*
- *Indépendance*
- *Probabilités conditionelles*
- *Variance et covariance*

3.1 Dérivée de Radon-Nikodym

Mesure σ-finie

Lorsque l'on travaille avec des mesures de probabilité, il est fréquent de vouloir exprimer une mesure en termes d'une autre à l'aide d'une densité. La dérivée de Radon-Nikodym permet justement de formuler cette idée de manière rigoureuse. Pour introduire la notion de dérivée de Radon-Nikodym, nous avons besoin de rappeler certaines propriétés fondamentales des mesures. L'une d'elles est la notion de mesure σ-finie.

Définition 3.1. Soit un espace mesuré (X, Σ_X, ν), la mesure ν est dite σ-finie s'il existe des ensembles $A_1, A_2, \cdots \in \Sigma_X$ tels que $\cup_{n\in\mathbb{N}} A_n = X$ où nous avons

$$\forall n \in \mathbb{N}, \quad \nu(A_n) < +\infty.$$

Exemple 3.2. *Voici trois exemples de mesure σ-finie :*

- *Une mesure de probabilité μ*
- *La mesure de Lebesgue sur $\mathbb{R}^d$*
- *La mesure de comptage (que l'on verra dans quelques instants)*

Mesure absolument continue

Une autre notion importante pour définir la dérivée de Radon-Nikodym est l'absolue continuité d'une mesure par rapport à une autre qui est définie dans la définition suivante.

Définition 3.3. Soit un espace mesurable (X, Σ_X) où nous avons deux mesures $\nu_1 : \Sigma_X \to [0, +\infty]$ et $\nu_2 : \Sigma_X \to [0, +\infty]$. On dit que la mesure ν_1 *est absolument continue par rapport à* ν_2 (que l'on note $\nu_1 \ll \nu_2$) si

$$\forall A \in \Sigma_X, \quad \nu_2(A) = 0 \implies \nu_1(A) = 0.$$

Cela signifie intuitivement que si une mesure attribue une masse nulle à un ensemble, alors l'autre mesure en fait de même.

Densité

Nous sommes maintenant prêt à introduire la dérivée de Radon-Nikodym, *i.e.*, la densité d'une mesure de probabilité μ par rapport à une autre ν.

Définition 3.4. Soit un espace mesurable (Ω, Σ_Ω) où nous avons une mesure de probabilité $\mu : \Sigma_\Omega \to [0,1]$ et une mesure σ-finie $\nu : \Sigma_\Omega \to [0, +\infty]$ tels que $\mu \ll \nu$. La mesure de probabilité μ possède une *densité* lorsque nous avons une fonction mesurable positive notée $^{d\mu}/_{d\nu}$ telle que

$$\forall A \in \Sigma_\Omega, \quad \mathop{\mathbb{P}}_{\omega\sim\mu}[\omega \in A] = \mathop{\mathbb{E}}_{\omega\sim\mu}\Big[\mathbb{1}[\omega \in A]\Big] = \int_\Omega \left(\mathbb{1}[\omega \in A]\frac{d\mu}{d\nu}(\omega)\right) d\nu(\omega),$$

où $\frac{d\mu}{d\nu}$ est appelé densité de μ par rapport à ν ou encore dérivée de Radon-Nikodym de μ par rapport à ν.

L'existence et l'unicité de cette fonction sont garanties par le théorème de Radon-Nikodym (non abordé ici). Ce résultat est fondamental en théorie des probabilités et permet de justifier la notion de densité de probabilité.

Fonction de masse

Nous pouvons maintenant définir des espérances sur des espaces finis ou dénombrable. Pour cela il faut définir la dérivée de Radon-Nikodym d'une mesure de probabilité μ en fonction de la mesure de comptage ν définie dans la définition suivante.

Définition 3.5 (Mesure de comptage)**.** Soit l'espace probabilisable $(X, 2^X)$ où Ω est fini ou dénombrable. La mesure de comptage $\nu : 2^X \to [0, +\infty]$ est définie telle que

$$\forall A \in \Sigma_X, \quad \nu(A) = \begin{cases} |A| & \text{si } A \text{ est fini,} \\ +\infty & \text{si } A \text{ est infini.} \end{cases}$$

En d'autres mots, la mesure de comptage est un cas particulier de mesure qui attribue à chaque ensemble son nombre d'éléments.
Dans le cas où Ω est fini ou dénombrable, la dérivée de Radon-Nikodym devient ce que l'on appelle la fonction de masse.

Définition 3.6 (Fonction de masse)**.** Soit (Ω, Σ_Ω) un espace mesurable avec un univers Ω fini ou dénombrable, où ν est la mesure de comptage et μ une mesure de probabilité tel que $\mu \ll \nu$. Alors $\frac{d\mu}{d\nu}(\omega)$ est appelée fonction de masse.

De plus, nous pouvons prouver la proposition suivante.

Proposition 3.7. Soit (Ω, Σ_Ω) un espace mesurable avec un univers Ω fini ou dénombrable, où ν est la mesure de comptage et μ une mesure de probabilité tel que $\mu \ll \nu$. Alors nous avons

$$\frac{d\mu}{d\nu}(\omega) = \mu(\{\omega\}).$$

Notation. Plus tard, nous notons $\mu(\omega) = \mu(\{\omega\})$.

Nous avons maintenant une expression plus explicite de l'espérance lorsque Ω est fini ou dénombrable.

Corollaire 3.8 (Espérance sur Ω fini ou dénombrable)**.** Soit (Ω, Σ_Ω) un espace mesurable avec un univers Ω fini ou dénombrable, où ν est la mesure de comptage et μ une mesure de probabilité tel que $\mu \ll \nu$. Nous avons

$$\mathop{\mathbb{E}}_{\omega\sim\mu}[f(\omega)] = \sum_{\omega\in X} \mu(\omega) f(\omega).$$

En effet, comme nous pouvons le voir, l'espérance peut s'écrire comme une somme de $\omega \in \Omega$ pondérée par la valeur de la fonction $f(\omega)$ et la valeur de la fonction de masse $\mu(\omega)$.
L'avantage de la fonction de masse est que l'on également définir plus intuitivement le passage de la mesure de probabilité μ_1 à μ_2 (que l'on appelle changement de mesure).

Corollaire 3.9 (Changement de mesure)**.** Soit (Ω, Σ_Ω) un espace mesurable où Ω est fini ou dénombrable. Soient ν la mesure de comptage, μ_1 une mesure de probabilité où $\nu \ll \mu_1 \ll \nu$, et μ_2 une mesure de probabilité où $\mu_2 \ll \nu$. Nous avons

$$\mathop{\mathbb{E}}_{\omega\sim\mu_1}\left[\frac{\mu_2(\omega)}{\mu_1(\omega)}f(\omega)\right] = \mathop{\mathbb{E}}_{\omega\sim\mu_2}\Big[f(\omega)\Big].$$

Fonction de densité

Pour une mesure de probabilité définie sur un sous-ensemble de $\mathbb{R}^d$, il est souvent utile d'introduire une fonction de densité qui permet d'exprimer cette mesure en fonction de la mesure de Lebesgue.

Définition 3.10 (Fonction de densité)**.** Soient (Ω, Σ_Ω) un espace mesurable où $\Omega \subseteq \mathbb{R}^d$, où ν est la mesure de Lebesgue et μ une mesure de probabilité. Alors $\frac{d\mu}{d\nu}(\omega)$ est appelée fonction de densité.

En d'autres termes, la fonction de densité permet de représenter la mesure μ en fonction de la mesure de Lebesgue. Cela permet de simplifier les calculs d'intégration, notamment pour l'espérance.

Notation. Plus tard, nous notons $\mu(\omega) = \frac{d\mu}{d\nu}(\omega)$.

Une des conséquences directes de cette définition est que l'espérance mathématique d'une fonction f sous une mesure μ peut être calculée sous forme d'une intégrale par rapport à la mesure de Lebesgue.

Corollaire 3.11 (Espérance sur $\Omega \subseteq \mathbb{R}^d$)**.** Soient (Ω, Σ_Ω) un espace mesurable avec $\Omega \subseteq \mathbb{R}^d$, où ν est la mesure de Lebesgue et μ est une mesure de probabilité telle que $\mu \ll \nu$. Nous avons

$$\begin{aligned}\mathop{\mathbb{E}}_{\omega\sim\mu}[f(\omega)] &= \int_X \mu(\omega) f(\omega) d\nu(\omega) \\ &= \int_X \mu(\omega) f(\omega) d\omega.\end{aligned}$$

Une autre propriété importante des densités de probabilité est qu'elles permettent de changer de mesure de probabilité.

Corollaire 3.12 (Changement de mesure)**.** Soit (Ω, Σ_Ω) un espace mesurable où $\Omega \subseteq \mathbb{R}^d$. Soient ν la mesure de Lebesgue, μ_1 une mesure de probabilité où $\nu \ll \mu_1 \ll \nu$, et μ_2 une mesure de probabilité où $\mu_2 \ll \nu$. Nous avons

$$\mathop{\mathbb{E}}_{\omega\sim\mu_1}\left[\frac{\mu_2(\omega)}{\mu_1(\omega)}f(\omega)\right] = \mathop{\mathbb{E}}_{\omega\sim\mu_2}\Big[f(\omega)\Big].$$

Changement de mesure (le cas général)

Dans le cas général, lorsque nous avons une mesure de référence (comme la mesure de comptage ou la mesure de Lebesgue), nous prenons le raccourcie de notation $\frac{d\mu_1}{d\nu}(\omega) = \mu_1(\omega)$ et $\frac{d\mu_2}{d\nu}(\omega) = \mu_2(\omega)$ et nous avons le changement de mesure suivant (généralisant les deux précédants).

Corollaire 3.13. Soit (Ω, Σ_Ω) un espace mesurable. Soient ν une mesure de référence sur Ω, μ_1 une mesure de probabilité où $\nu \ll \mu_1 \ll \nu$, et μ_2 une mesure de probabilité où $\mu_2 \ll \nu$. Nous avons

$$\mathop{\mathbb{E}}_{\omega\sim\mu_1}\left[\frac{\mu_2(\omega)}{\mu_1(\omega)}f(\omega)\right] = \mathop{\mathbb{E}}_{\omega\sim\mu_2}\Big[f(\omega)\Big].$$

Dans le cas où nous n'avons pas de mesure de références, nous avons le résultat suivant.

Corollaire 3.14. Soient (Ω, Σ_Ω) un espace mesuré où μ_1 et μ_2 sont deux mesures de probabilité telles que $\mu_2 \ll \mu_1$. Nous avons

$$\mathop{\mathbb{E}}_{\omega\sim\mu_1}\left[\frac{d\mu_2}{d\mu_1}(\omega)f(\omega)\right] = \mathop{\mathbb{E}}_{\omega\sim\mu_2}\Big[f(\omega)\Big].$$

3.2 Variable aléatoire

Définition

Définition 3.15. Soit $(\Omega, \Sigma_\Omega, \mu)$ un espace probabilisé, et (E, Σ_E) est un espace mesurable. Une *variable aléatoire* est une fonction mesurable $f : \Omega \to E$.

Exemple. *Voici trois exemples de variable aléatoire :*

- *Si* $(E, \Sigma_E) = (\mathbb{R}, \textit{Borel}(\mathbb{R}))$ *alors* f *est une* variable aléatoire réelle.
- *Si* E *est un ensemble fini ou dénombrable alors* f *est une* variable aléatoire discrète.
- *La fonction* $f : \omega \mapsto \omega$ *(où* $E = \Omega$*) est une variable aléatoire.*

Voici un exemple concret de variable aléatoire réelle.

Exemple. *Soit* $\Omega = \{$⚀☼, ⚁☼, ⚂☼, ⚃☼, ⚄☼, ⚅☼, ⚀☁, ⚁☁, ⚂☁, ⚃☁, ⚄☁, ⚅☁$\}$.
La variable aléatoire $f : \Omega \to \mathbb{R}$ *suivante donne la valeur du dé :*

$$\begin{aligned}
\text{⚀☼} &\mapsto 1\\
\text{⚁☼} &\mapsto 2\\
\text{⚂☼} &\mapsto 3\\
\text{⚃☼} &\mapsto 4\\
\text{⚄☼} &\mapsto 5\\
\text{⚅☼} &\mapsto 6\\
\text{⚀☁} &\mapsto 1\\
\text{⚁☁} &\mapsto 2\\
\text{⚂☁} &\mapsto 3\\
\text{⚃☁} &\mapsto 4\\
\text{⚄☁} &\mapsto 5\\
\text{⚅☁} &\mapsto 6
\end{aligned}$$

Grâce à cette variable aléatoire, nous avons par exemple :

$$\mathop{\mathbb{P}}_{\omega\sim\mu}\Big[f(\omega) \in \{1,2\}\Big] = \mathop{\mathbb{P}}_{\omega\sim\mu}\Big[\omega \in \{\text{⚀☼}, \text{⚁☼}, \text{⚀☁}, \text{⚁☁}\}\Big].$$

Etant donné que ω peut être vu comme une variable aléatoire, nous allons souvent utiliser la notation suivante pour définir une variable aléatoire.

Définition 3.16. On dit que $\omega \sim \mu$, qui se lit "ω suit la loi / distribution μ sur Ω", si nous avons un espace probabilisé $(\Omega, \Sigma_\Omega, \mu)$.

Cette notation est (je l'espère) intuitive : on peut voir ω comme le fruit d'une réalisation d'une expérience aléatoire. Autrement dit, si on imagine μ comme un "chapeau magique", alors ω est un résultat sorti de ce "chapeau magique".

Notation. Lorsque nous écrirons $\omega \sim \mu$, où μ est une distribution sur Ω, nous considérerons que tout se passe bien pour définir la tribu Σ_Ω.

Lorsque nous avons une fonction (mesurable) f, nous pouvons définir une autre mesure de probabilité μ_f qui correspond à la loi de probabilité pour f.

Définition 3.17. Soit une variable aléatoire $\omega \sim \mu$ où μ est une mesure de probabilité sur Ω et $f : \Omega \to \Omega'$ une variable aléatoire. Nous pouvons définir la mesure de probabilité μ_f sur Ω' associée à f de la façon suivante :

$$\begin{aligned}
\mu_f(A) &:= \mathop{\mathbb{P}}_{f(\omega)\sim\mu_f}[f(\omega) \in A]\\
&:= \mathop{\mathbb{P}}_{\omega\sim\mu}\left[\omega \in f^{-1}(A)\right].
\end{aligned}$$

Ainsi, nous avons le corollaire suivant.

Corollaire 3.18. Si nous avons $\omega \sim \mu$ où μ est une mesure de probabilité sur Ω et $f : \Omega \to \Omega'$ une variable aléatoire. alors nous avons $f(\omega) \sim \mu_f$.

Théorème de transfert

Le théorème de transfert permet de transférer l'intégration de Ω vers Ω'. En effet, nous avons le résultat suivant.

Proposition 3.19 (théorème de transfert)**.** Soient $\omega \sim \mu$, $f(\omega) \sim \mu_f$ et $g : \Omega' \to \mathbb{R}$ une fonction mesurable alors nous avons

$$\mathop{\mathbb{E}}_{\omega\sim\mu}[g(f(\omega))] = \mathop{\mathbb{E}}_{f(\omega)\sim\mu_f}[g(f(\omega))].$$

3.3 Indépendance et théorèmes de Fubini

Nous allons maintenant voir la notion d'indépendance qui est fondamentale en théorie des probabilités et statistiques. Pour cela, nous allons (encore) avoir besoin d'une notion venant de la théorie de la mesure : les tribus et les mesures produits.

Tribus et mesures produits

Afin de définir une mesure produit, nous avons besoin de définir la tribu produit.

Définition 3.20. Soient $(\Omega_1, \Sigma_{\Omega_1})$ et $(\Omega_2, \Sigma_{\Omega_2})$ deux espaces probabilisables, la *tribu produit* est définie par

$$\Sigma_{\Omega_1}\otimes\Sigma_{\Omega_2} := \sigma\left(\{A_1 \times A_2 | A_1 \in \Sigma_{\Omega_1}, A_2 \in \Sigma_{\Omega_2}\}\right),$$

où $\sigma(A)$ est la tribu engendrée par l'ensemble A.

Nous somme maintenant prêts à définir la mesure produit.

Définition 3.21. Une *mesure produit* $\mu_1\otimes\mu_2$ est une mesure sur l'espace probabilisable $(\Omega_1\times\Omega_2, \Sigma_{\Omega_1}\otimes\Sigma_{\Omega_2})$. Elle est définie par

$$\mu_1\otimes\mu_2(A_1 \times A_2) := \mu_1(A_1)\mu_2(A_2).$$

Variables aléatoires indépendantes

Grâce aux mesures produits, nous sommes (enfin) capable de définir l'indépendance dans les variables aléatoires.

Définition 3.22. Soient $\omega = (\omega_1, \omega_2) \sim \mu$, on dit que ω_1 et ω_2 sont indépendants si $\mu = \mu_1 \otimes \mu_2$ où $\omega_1 \sim \mu_1$ et $\omega_2 \sim \mu_2$.

Ainsi, grâce à la définition de l'indépendance et la définition d'une mesure, nous avons le corollaire suivant.

Corollaire 3.23. Si ω_1 et ω_2 sont indépendant si et seulement si

$$\mathop{\mathbb{P}}_{(\omega_1,\omega_2)\sim\mu}\left[\omega_1 \in A_1, \omega_2 \in A_2\right] = \mathop{\mathbb{P}}_{\omega_1\sim\mu_1}[\omega_1 \in A_1]\mathop{\mathbb{P}}_{\omega_2\sim\mu_2}[\omega_2 \in A_2]$$

Voici un exemple de variables aléatoires dépendantes.

Exemple. *Soient*

- ω_1 *le résultat d'un dé à 6 faces,*
- $\omega_2 = \omega_1 + 1$.

Alors, nous avons :

$$\mathop{\mathbb{P}}_{(\omega_1,\omega_2)\sim\mu}[\omega_1 = 1 \text{ et } \omega_2 = 3] = 0,$$
$$\mathop{\mathbb{P}}_{(\omega_1,\omega_2)\sim\mu}[\omega_1 = 1] = \frac{1}{6},$$
$$\text{et } \mathop{\mathbb{P}}_{(\omega_1,\omega_2)\sim\mu}[\omega_2 = 3] = \frac{1}{6},$$

Voici un exemple de variables aléatoires indépendantes.

Exemple. *Soient*

- ω_1 *le résultat d'un lancer de dé à 6 faces,*
- ω_2 *le résultat d'un deuxième lancer de dé à 6 faces.*

Théorème de Fubini-Tonelli

L'avantage d'avoir des variables indépendantes est que l'on peut ensuite intervertir leurs espérances. Ceci peut être effectuée grâce à l'un des deux théorèmes de Fubini.

Théorème 3.24. Soient $(\Omega_1, \Sigma_{\Omega_1}, \mu_1)$ et $(\Omega_2, \Sigma_{\Omega_2}, \mu_2)$ deux espaces probabilisés. Soient $(\Omega_1 \times \Omega_2, \Sigma_{\Omega_1} \otimes \Sigma_{\Omega_2}, \mu_1 \otimes \mu_2)$ un espace probabilisable muni de la mesure produit et $f : \Omega_1 \times \Omega_2 \to [0, +\infty]$ est une fonction mesurable alors

$$\begin{aligned} \mathop{\mathbb{E}}_{(\omega_1,\omega_2)\sim\mu_1\otimes\mu_2}[f(\omega_1,\omega_2)] &= \mathop{\mathbb{E}}_{\omega_1\sim\mu_1}\left[\mathop{\mathbb{E}}_{\omega_2\sim\mu_2}[f(\omega_1,\omega_2)]\right] \\ &= \mathop{\mathbb{E}}_{\omega_2\sim\mu_2}\left[\mathop{\mathbb{E}}_{\omega_1\sim\mu_1}[f(\omega_1,\omega_2)]\right]. \end{aligned}$$

Théorème de Fubini-Lebesgue

Théorème 3.25. Soient $(\Omega_1, \Sigma_{\Omega_1}, \mu_1)$ et $(\Omega_2, \Sigma_{\Omega_2}, \mu_2)$ deux espaces probabilisés. Soient $(\Omega_1 \times \Omega_2, \Sigma_{\Omega_1} \otimes \Sigma_{\Omega_2}, \mu_1 \otimes \mu_2)$ un espace probabilisable muni de la mesure produit. Si $f : \Omega_1 \times \Omega_2 \to \mathbb{R}$ est intégrable, c'est à dire si

$$\mathop{\mathbb{E}}_{(\omega_1,\omega_2)\sim\mu_1\otimes\mu_2}\left[|f(\omega_1,\omega_2)|\right] < +\infty$$

alors nous avons

$$\begin{aligned} \mathop{\mathbb{E}}_{(\omega_1,\omega_2)\sim\mu_1\otimes\mu_2}[f(\omega_1,\omega_2)] &= \mathop{\mathbb{E}}_{\omega_1\sim\mu_1}\left[\mathop{\mathbb{E}}_{\omega_2\sim\mu_2}[f(\omega_1,\omega_2)]\right] \\ &= \mathop{\mathbb{E}}_{\omega_2\sim\mu_2}\left[\mathop{\mathbb{E}}_{\omega_1\sim\mu_1}[f(\omega_1,\omega_2)]\right]. \end{aligned}$$

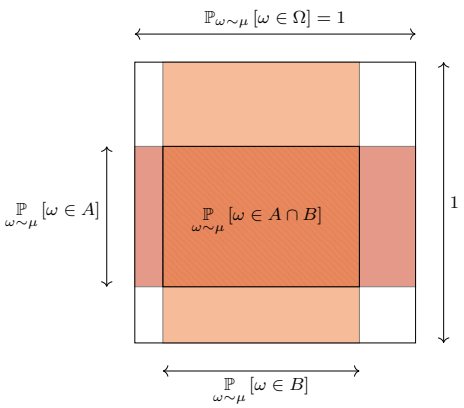
Événements indépendants

Enfin, il existe une notion d'indépendance qui n'est pas sur les variables aléatoires mais sur les événements.

Définition 3.26. Soit $\omega \sim \mu$ où μ est une distribution sur Ω, *deux événements* $A \in \Sigma_\Omega$ et $B \in \Sigma_\Omega$ sont *indépendants* si et seulement si

$$\mathop{\mathbb{P}}_{\omega\sim\mu}[\omega \in A \cap B] = \mathop{\mathbb{P}}_{\omega\sim\mu}[\omega \in A] \cdot \mathop{\mathbb{P}}_{\omega\sim\mu}[\omega \in B].$$

Exemple. *Nous pouvons (facilement) représenter graphiquement deux événements indépendants.*



3.4 Probabilités conditionelles

La probabilité conditionelle correspond au calcul de nouvelles probabilités, en supposant que l'on apprenne qu'un événement B s'est réalisé ou va se réaliser. On suppose que l'événement B est compatible avec notre ensemble de mondes possibles. Dit autrement, l'événement B est probable, c'est-à-dire que $\mathbb{P}_{\omega\sim\mu}[\omega \in B] > 0$.

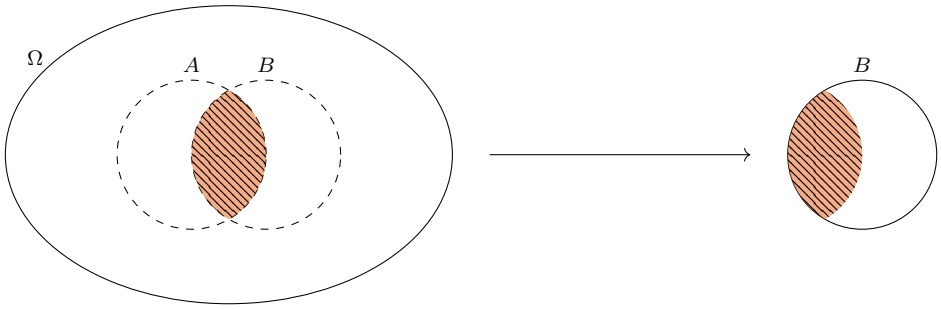
Définition 3.27. Soit $(\Omega, \Sigma_\Omega, \mu)$ un espace probabilisé, *i.e.*, $\omega \sim \mu$ où μ est une mesure de probabilité sur Ω. Soit $B \in \Sigma_\Omega$ avec $\mathbb{P}_{\omega\sim\mu}[\omega \in B] > 0$, on note la *probabilité (conditionelle)* de $A \in \Sigma_\Omega$ étant donné B par

$$\mathop{\mathbb{P}}_{\omega\sim\mu}[\omega \in A \mid B] := \frac{\mathbb{P}_{\omega\sim\mu}[\omega \in A \cap B]}{\mathbb{P}_{\omega\sim\mu}[\omega \in B]}.$$

La notation $\mathbb{P}_{x\sim\mu}[x \in A \mid B]$, bien que classique, est un peu ambiguë, car $A|B$ ne désigne pas un événement !

Comme le montre le schéma ci-dessous, cela revient à considérer que le nouvel ensemble Ω est maintenant B, et de normaliser la mesure de probabilité par $\mathbb{P}_{\omega\sim\mu}[\omega \in B]$.

Exemple.



La probabilité conditionelle permet en réalité de définir une nouvelle mesure de probabilité !

Proposition 3.28. Soit $(\Omega, \Sigma_\Omega, \mu)$ un espace probabilisé, *i.e.*, $\omega \sim \mu$ où μ est une mesure de probabilité sur Ω. Soit $B \in \Sigma_\Omega$ avec $\mathbb{P}_{\omega\sim\mu}[\omega \in B] > 0$, alors notons

$$\mu_B(\cdot) := \mathop{\mathbb{P}}_{\omega\sim\mu}[\omega \in \cdot \mid B],$$

et $(\Omega, \Sigma_\Omega, \mu_B)$ est un espace probabilisé, *i.e.*, μ_B est une mesure de probabilité sur Ω.

Ainsi, nous pouvons réécrire la probabilité conditionelle de la façon suivante.

Corollaire 3.29. Soit $\omega \sim \mu$ où μ est une mesure de probabilité sur Ω. Soit $B \in \Sigma_\Omega$ avec $\mathbb{P}_{\omega\sim\mu}[\omega \in B] > 0$, nous avons

$$\forall A \in \Sigma_\Omega, \quad \mathop{\mathbb{P}}_{\omega\sim\mu_B}[\omega \in A] = \frac{\mathbb{P}_{\omega\sim\mu}[\omega \in A \cap B]}{\mathbb{P}_{\omega\sim\mu}[\omega \in B]}.$$

Propriétés

Les probabilités conditionelles nous permettent de dériver des propriétés assez naturelles (et pour certaine, très connues).

Proposition 3.30. Soit $\omega \sim \mu$ où μ est une mesure de probabilité sur Ω. Si $A \in \Sigma_\Omega$ et $B \in \Sigma_\Omega$ sont deux événements indépendants, alors

$$\mathop{\mathbb{P}}_{\omega\sim\mu_B}[\omega \in A] = \mathop{\mathbb{P}}_{\omega\sim\mu}[\omega \in A].$$

Proposition 3.31 (règle du produit)**.** Soit $\omega \sim \mu$ où μ est une mesure de probabilité sur Ω. Soit $A \in \Sigma_\Omega$ et $B \in \Sigma_\Omega$ avec $\mathbb{P}_{\omega\sim\mu}[\omega \in A] > 0$ et $\mathbb{P}_{\omega\sim\mu}[\omega \in B] > 0$, alors nous avons

$$\mathop{\mathbb{P}}_{\omega\sim\mu}[\omega \in A \cap B] = \mathop{\mathbb{P}}_{\omega\sim\mu_B}[\omega \in A] \cdot \mathop{\mathbb{P}}_{\omega\sim\mu}[\omega \in B] = \mathop{\mathbb{P}}_{\omega\sim\mu_A}[\omega \in B] \cdot \mathop{\mathbb{P}}_{\omega\sim\mu}[\omega \in A].$$

Théorème 3.32 (formule de Bayes)**.** Soit $\omega \sim \mu$ où μ est une mesure de probabilité sur Ω. Soit $A \in \Sigma_\Omega$ et $B \in \Sigma_\Omega$ avec $\mathbb{P}_{\omega\sim\mu}[\omega \in A] > 0$ et $\mathbb{P}_{\omega\sim\mu}[\omega \in B] > 0$. Nous avons

$$\mathop{\mathbb{P}}_{\omega\sim\mu_B}[\omega \in A] = \frac{\mathbb{P}_{\omega\sim\mu}[\omega \in A]}{\mathbb{P}_{\omega\sim\mu}[\omega \in B]} \mathop{\mathbb{P}}_{\omega\sim\mu_A}[\omega \in B]$$

$$\iff \quad \mathop{\mathbb{P}}_{\omega\sim\mu_B}[\omega \in A] \cdot \mathop{\mathbb{P}}_{\omega\sim\mu}[\omega \in B] = \mathop{\mathbb{P}}_{\omega\sim\mu_A}[\omega \in B] \cdot \mathop{\mathbb{P}}_{\omega\sim\mu}[\omega \in A].$$

3.5 Variance

L'objectif est de mesurer l'écart par rapport à l'espérance. La variance est définie comme l'espérance des écarts au carré par rapport à la moyenne.

Définition

La variance de $f(\omega)$ où $\omega \sim \mu$ et $f : \Omega \to \mathbb{R}$ n'est définie que lorsque $f(\omega)^2$ est intégrable.

Définition 3.33. $f(\omega)$ où $\omega \sim \mu$ admet un moment d'ordre 2 si $f(\omega)^2$ est intégrable, *i.e.*, si $\mathbb{E}_{\omega\sim\mu}\left[f(\omega)^2\right] < +\infty$.

Nous pouvons maintenant définir la variance d'une variable aléatoire $f(\omega)$ où $\omega \sim \mu$.

Définition 3.34. Si $f(\omega)$ admet un moment d'ordre 2, la variance de f est

$$\mathop{\mathbb{V}}_{\omega\sim\mu}[f(\omega)] := \mathop{\mathbb{E}}_{\omega\sim\mu}\left[\left(f(\omega) - \mathop{\mathbb{E}}_{\omega'\sim\mu}[f(\omega')]\right)^2\right].$$

De plus, la variance est liée à une autre notion : l'écart-type.

Définition 3.35. L'écart-type est $\sqrt{\mathop{\mathbb{V}}_{\omega\sim\mu}[f(\omega)]}$.

Propriétés

Nous pouvons également obtenir les deux propriétés suivantes.

Proposition 3.36 (formule de König-Huygens)**.** $\mathbb{V}_{\omega\sim\mu}[f(\omega)] = \mathbb{E}_{\omega\sim\mu}\left[f(\omega)^2\right] - (\mathbb{E}_{\omega\sim\mu}[f(\omega)])^2$.

Proposition 3.37. $\mathbb{V}_{\omega\sim\mu}[af(\omega) + b] = a^2\,\mathbb{V}_{\omega\sim\mu}[f(\omega)]$.

Covariance

On cherche à mesurer si deux variables $f_1(\omega_1)$ et $f_2(\omega_2)$ sont corrélées, c'est-à-dire si un écart de $f_1(\omega_1)$ à sa moyenne, se traduit aussi par un écart de $f_2(\omega_2)$ à sa moyenne. Pour mesurer la corrélation, on définit la covariance comme l'espérance du produit des écarts.

Définition 3.38 (covariance)**.** La covariance est

$$\mathop{\mathrm{cov}}_{(\omega_1,\omega_2)\sim\mu}(f_1(\omega_1), f_2(\omega_2)) := \left[\mathop{\mathbb{E}}_{(\omega_1,\omega_2)\sim\mu}\left(f_1(\omega_1) - \mathop{\mathbb{E}}_{(\omega_1',\omega_2')\sim\mu}\left[f_2(\omega_2')\right]\right)\left(f_2(\omega_2) - \mathop{\mathbb{E}}_{(\omega_1',\omega_2')\sim\mu}\left[f_2(\omega_2')\right]\right)\right].$$

Propriétés

Nous pouvons également obtenir les deux propriétés suivantes.

Proposition 3.39 (formule de König-Huygens)**.**

$$\underset{(\omega_1,\omega_2)\sim\mu}{\mathrm{cov}}(f_1(\omega_1), f_2(\omega_2)) = \underset{(\omega_1,\omega_2)\sim\mu}{\mathbb{E}}\Big[f_1(\omega_1)f_2(\omega_2)\Big] - \underset{(\omega_1,\omega_2)\sim\mu}{\mathbb{E}}\Big[f_1(\omega_1)\Big]\underset{(\omega_1,\omega_2)\sim\mu}{\mathbb{E}}\Big[f_2(\omega_2)\Big].$$

Définition 3.40.
$f_1(\omega_1)$ et $f_2(\omega_2)$ sont *corrélées* si $\mathrm{cov}_{(\omega_1,\omega_2)\sim\mu}(f_1(\omega_1), f_2(\omega_2)) \neq 0$.
$f_1(\omega_1)$ et $f_2(\omega_2)$ sont *corrélées positivement* si $\mathrm{cov}_{(\omega_1,\omega_2)\sim\mu}(f_1(\omega_1), f_2(\omega_2)) > 0$.
$f_1(\omega_1)$ et $f_2(\omega_2)$ sont *corrélées négativement* si $\mathrm{cov}_{(\omega_1,\omega_2)\sim\mu}(f_1(\omega_1), f_2(\omega_2)) < 0$.

Proposition 3.41. $f_1(\omega_1)$ et $f_2(\omega_2)$ indépendantes *implique* que $\mathrm{cov}_{(\omega_1,\omega_2)\sim\mu}(f_1(\omega_1), f_2(\omega_2)) = 0$.

Chapitre 4

Lois

4.1 Lois discrètes

Loi de Dirac

Nous allons tout d'abord voir la loi (ou distribution) de Dirac. C'est la distribution la plus simple qui correspond à ne pas avoir d'aléatoire.

Définition 4.1. ω suit une *loi de Dirac*, notée δ_a, en a, c'est-à-dire que $\omega \sim \delta_a$, si

$$\mathop{\mathbb{P}}_{\omega\sim\delta_a}[\omega = a] = 1.$$

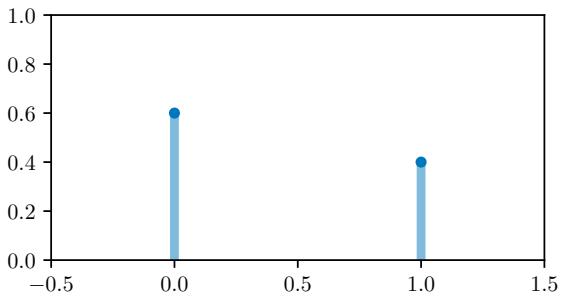
Loi de Bernoulli

Une loi de Bernoulli $\mathcal{B}(p)$ modélise le lancer d'une pièce pipée où il y a une probabilité p de faire pile, et $1-p$ de faire face. Pile est comptabilisé comme 1 (succès ✔ de faire pile), et face comme 0 (échec ✖ de faire pile).

Définition 4.2. Soit $p \in [0,1]$, ω suit une *loi de Bernoulli*, notée $\mathcal{B}(p)$, c'est-à-dire que $\omega \sim \mathcal{B}(p)$, si

$$\underbrace{\mathop{\mathbb{P}}_{\omega\sim\mathcal{B}(p)}[\omega = 1] = p}_{\text{✔ succès}} \quad \text{et} \quad \underbrace{\mathop{\mathbb{P}}_{\omega\sim\mathcal{B}(p)}[\omega = 0]}_{\text{✖ échec}} = 1-p.$$

Exemple. *Fonction de masse de la loi de Bernoulli* $\mathcal{B}(0.4)$.



Nous pouvons déduire les propriétés suivantes pour la loi de Bernoulli.

Proposition 4.3. Soit $\omega \sim \mathcal{B}(p)$ avec $p \in [0,1]$. Nous avons

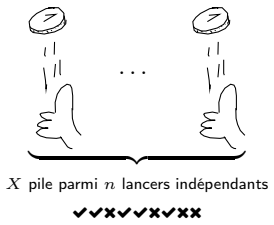
$$\mathop{\mathbb{E}}_{\omega\sim\mathcal{B}(p)}[\omega] = p.$$

Proposition 4.4. Soit $\omega \sim \mathcal{B}(p)$ avec $p \in [0,1]$. Nous avons

$$\mathop{\mathbb{V}}_{\omega\sim\mathcal{B}(p)}[\omega] = p(1-p).$$

Loi binomiale

Une loi binomiale $\mathcal{B}(n,p)$ correspond au nombre de succès ✔ sur n essais. Dit autrement, c'est le nombre de fois que l'on a obtenu pile avec une pièce, lancée n fois, et toujours avec probabilité p de faire pile sur un lancer.



Définition 4.5. Soit $p \in [0,1]$ et $n \in \mathbb{N}_*$, $\mathcal{B}(n,p)$ est la loi de $N(\omega_1, \dots, \omega_n) = \omega_1 + \cdots + \omega_n$ où les $\omega_i \sim \mathcal{B}(p)$ sont *i.i.d.*.

Remarque 4.6. *On voit que* $\mathcal{B}(1,p) = \mathcal{B}(p)$.

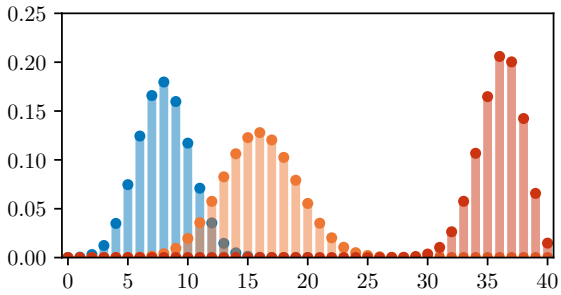
Les lancers sont bien indépendants. Hors de question, par exemple que pour le 5e lancer, on recopie la valeur obtenue au 4e lancer ; les 4e et 5e lancers sont indépendants, c'est-à-dire qu'il faut bien relancer la pièce. Aussi, le résultat du 4e lancer n'influence pas du tout le résultat du 5e lancer.
Nous pouvons également trouver l'expression de $\mathbb{P}_{N\sim\mathcal{B}(n,p)}[N=k]$.

Proposition 4.7. Si $N \sim \mathcal{B}(n,p)$ avec $p \in [0,1]$ et $n \in \mathbb{N}_*$, alors pour tout $k \in \{0, \dots, n\}$ nous avons

$$\mathop{\mathbb{P}}_{N\sim\mathcal{B}(n,p)}[N=k] = \overbrace{\binom{n}{k}}^{\text{Façons de placer } k \text{ ✔ dans } n \text{ essais}} \cdot \underbrace{p^k \cdot (1-p)^{n-k}}_{\text{proba. d'avoir } k \text{ ✔ et } n-k \text{ ✖}}$$

Du coup, $p^k(1-p)^{n-k}$ est la probabilité d'avoir k ✔ et $n-k$ ✖. $\binom{n}{k}$ est le nombre de sous-ensemble de cardinal k, c'est-à-dire le nombre de façons de placer ces k piles.

Exemple. *Fonctions de masse des lois binomiales* $\mathcal{B}(20, 0.4)$, $\mathcal{B}(40, 0.4)$, *et* $\mathcal{B}(40, 0.9)$.



Nous pouvons dériver les propriétés suivantes pour la loi binomiale.

Proposition 4.8. Soit $N \sim \mathcal{B}(n,p)$ avec $p \in [0,1]$ et $n \in \mathbb{N}_*$. Nous avons

$$\mathop{\mathbb{E}}_{N\sim\mathcal{B}(n,p)}[N] = np.$$

Proposition 4.9. Soit $N \sim \mathcal{B}(n,p)$ avec $p \in [0,1]$ et $n \in \mathbb{N}_*$. Nous avons

$$\mathop{\mathbb{V}}_{N\sim\mathcal{B}(n,p)}[N] = np(1-p).$$

Loi géométrique

La loi géométrique correspond au *nombre d'essais jusqu'à un succès* (essai avec succès compris). Par exemple, elle permet de calculer le nombre de lancers de pièce jusqu'à avoir pile (où nous avons par exemple 7 échecs ✖✖✖✖✖✖✖ suivis d'un succès ✔).

Définition 4.10. Soit $p \in [0,1]$, $\mathcal{G}(p)$ est la loi de $T(\omega_1, \dots) = \min(\{k \in \mathbb{N}^* \mid \omega_k = 1\})$ où $\omega_1, \omega_2, \dots \sim \mathcal{B}(p)$.

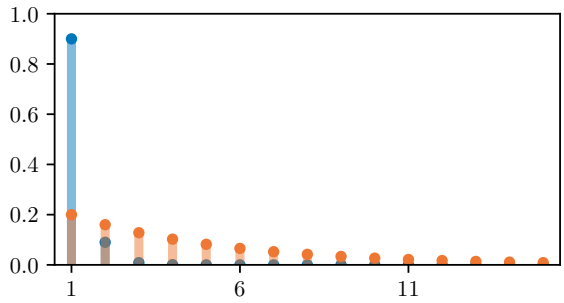
Nous pouvons en déduire l'expression suivante.

Proposition 4.11. Si $T \sim \mathcal{G}(p)$ avec $p \in [0,1]$, alors nous avons

$$\forall k \in \mathbb{N}^*, \quad \mathop{\mathbb{P}}_{T\sim\mathcal{G}(p)}[T = k] = \mathop{\mathbb{P}}_{\omega_1,\dots\sim\mathcal{B}(p)}[T(\omega_1, \dots) = k] = \underbrace{(1-p)^{k-1}p}_{\text{proba. d'avoir } k \text{ ✔ puis un ✖}},$$

où $T(\omega_1, \dots) = \min(\{k \in \mathbb{N}^* \mid \omega_k = 1\})$.

Exemple. *Fonctions de masse des lois géométrique* $\mathcal{G}(0.9)$, $\mathcal{G}(0.2)$.



Nous avons plusieurs propriétés que nous pouvons établir.

Proposition 4.12. Soit $T \sim \mathcal{G}(p)$ avec $p \in [0,1]$. Nous avons

$$\mathop{\mathbb{E}}_{T\sim\mathcal{G}(p)}[T] = \sum_{k\in\mathbb{N}^*} \mathop{\mathbb{P}}_{T\sim\mathcal{G}(p)}[T = k] \cdot k = \frac{1}{p}.$$

Proposition 4.13. Soit $T \sim \mathcal{G}(p)$ avec $p \in [0,1]$. Nous avons

$$\mathop{\mathbb{V}}_{T\sim\mathcal{G}(p)}[T] = \frac{(1-p)}{p^2}.$$

Proposition 4.14. Soit $T \sim \mathcal{G}(p)$ avec $p \in [0,1]$. Nous avons

$$\forall k \in \mathbb{N}^*, \quad \mathop{\mathbb{P}}_{T\sim\mathcal{G}(p)}[T > k] = (1-p)^k.$$

Proposition 4.15 (sans mémoire). Soit $T \sim \mathcal{G}(p)$ avec $p \in [0,1]$. Nous avons

$$\forall s, t \in \mathbb{N}, \quad \mathop{\mathbb{P}}_{T\sim\mathcal{G}(p)}[T > s + t \mid T > s] = \mathop{\mathbb{P}}_{T\sim\mathcal{G}(p)}[T > t].$$

Loi de Poisson

Si, dans la loi binomiale, p est très petit, et n grand, et en posant $\lambda = pn \in \mathbb{R}$, nous obtenons alors la loi de Poisson. En effet, nous avons

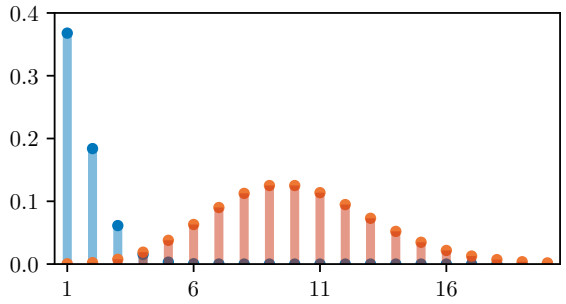
$$\begin{aligned}\mathop{\mathbb{P}}_{N\sim\mathcal{B}(n,p)}[N = k] = \binom{n}{k} p^k (1-p)^{n-k} &= \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \frac{n!(1-\frac{\lambda}{n})^{-k}}{(n-k)!n^k} \xrightarrow{n\to+\infty} \frac{\lambda^k}{k!} e^{-\lambda}.\end{aligned}$$

Définition 4.16. Si $N \sim \mathcal{P}(\lambda)$ avec $\lambda \in \mathbb{R}$, alors nous avons

$$\mathop{\mathbb{P}}_{N\sim\mathcal{P}(\lambda)}[N = k] = \frac{\lambda^k}{k!} e^{-\lambda}.$$

La loi de Poisson est utilisée pour modéliser des événements rares.

Exemple. *Fonctions de masse des lois de Poisson* $\mathcal{P}(1)$ *et* $\mathcal{P}(10)$.



Nous pouvons déterminer l'espérance et la variance d'une variable aléatoire N suivant une loi de Poisson.

Proposition 4.17. Soit $N \sim \mathcal{P}(\lambda)$ avec $\lambda \in \mathbb{R}$. Nous avons

$$\begin{aligned}\mathop{\mathbb{E}}_{N\sim\mathcal{P}(\lambda)}[N] &= \sum_{k\in\mathbb{N}^*} k e^{\lambda} \frac{\lambda^k}{k!} \\ &= \lambda e^{-\lambda} e^{\lambda} \\ &= \lambda.\end{aligned}$$

Proposition 4.18. Soit $N \sim \mathcal{P}(\lambda)$ avec $\lambda \in \mathbb{R}$. Nous avons

$$\mathop{\mathbb{V}}_{N\sim\mathcal{P}(\lambda)}[N] = \lambda.$$

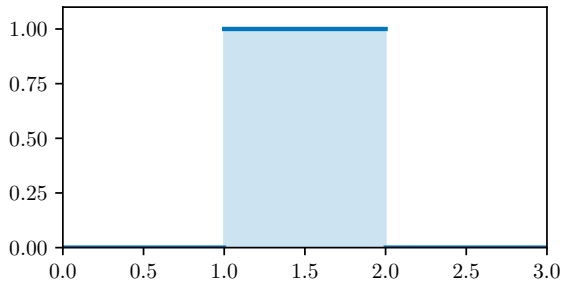
4.2 Lois continues (à densité)

Lois uniformes

Voici l'une des lois continues les plus simples que nous avons rencontrées dans le Chapitre 1.

Définition 4.19. Soit $a < b$ avec $a, b \in \mathbb{R}$. La *loi uniforme* $\mathcal{U}(a, b)$ est de densité $\mathcal{U}_{a,b}(x) = \frac{1}{b-a}\mathbb{1}[x \in [a, b]]$.

Exemple. *Fonction de densité* $\mathcal{U}_{1,2}$ *de la loi uniforme* $\mathcal{U}(1, 2)$.

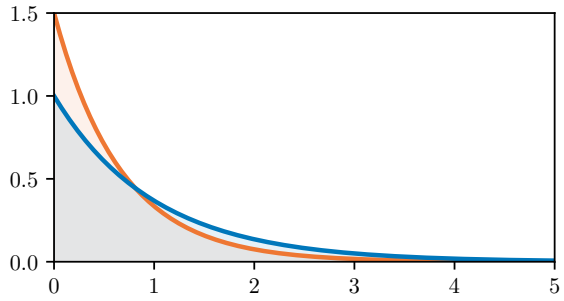


Loi exponentielle

Les lois exponentielles modélisent des *temps d'attente continus jusqu'à un succès*. Elles sont le pendant continu des lois géométriques.

Définition 4.20. La *loi exponentielle* $\mathcal{E}(\lambda)$ avec $\lambda \in \mathbb{R}$ est de densité $\mathcal{E}_\lambda(x) = \lambda e^{-\lambda \cdot x} \mathbb{1}\left[x \in \mathbb{R}^+\right]$.

Exemple. *Fonctions de densité $\mathcal{E}_\lambda$ des lois exponentielles $\mathcal{E}(1)$ et $\mathcal{E}(1.5)$.*



Voici l'espérance et la variance d'une variable aléatoire $T \sim \mathcal{E}(\lambda)$ suivant une loi exponentielle.

Proposition 4.21. Soit $T \sim \mathcal{E}(\lambda)$ avec $\lambda \in \mathbb{R}$. Nous avons

$$\mathop{\mathbb{E}}_{T\sim\mathcal{E}(\lambda)}[T] = \int_0^{+\infty} x\lambda e^{-\lambda x} dx = \frac{1}{\lambda}$$

Proposition 4.22. Soit $T \sim \mathcal{E}(\lambda)$ avec $\lambda \in \mathbb{R}$. Nous avons

$$\mathop{\mathbb{V}}_{T\sim\mathcal{E}(\lambda)}[T] = \frac{1}{\lambda^2}$$

Voici deux autres propriétés (similaires à celles de la loi géométrique).

Proposition 4.23. Soit $T \sim \mathcal{E}(\lambda)$ avec $\lambda \in \mathbb{R}$. Nous avons

$$\forall s \in \mathbb{R}^+, \quad \mathop{\mathbb{P}}_{T\sim\mathcal{E}(\lambda)}[T > s] = e^{-\lambda s}.$$

Proposition 4.24 (sans mémoire)**.** Soit $T \sim \mathcal{E}(\lambda)$ avec $\lambda \in \mathbb{R}$. Nous avons

$$\forall s, t \in \mathbb{R}^+, \quad \mathop{\mathbb{P}}_{T\sim\mathcal{E}(\lambda)}[T > s + t \mid T > s] = \mathop{\mathbb{P}}_{T\sim\mathcal{E}(\lambda)}[T > t].$$

Lois normales

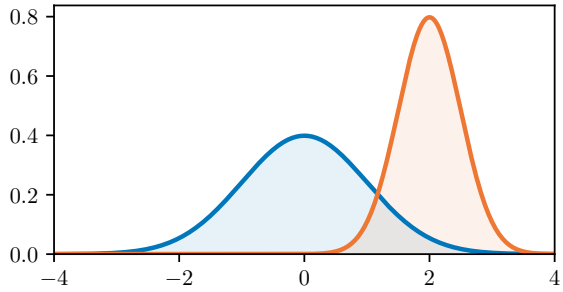
Voici l'une des lois les plus connues des probabilités/statistiques : la loi normale.

Définition 4.25. Soit $\mu \in \mathbb{R}$ et $\sigma > 0$, la *loi normale* $\mathcal{N}(\mu, \sigma^2)$ est de densité

$$\mathcal{N}_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}.$$

Définition 4.26. La *loi normale centrée réduite* est la loi $\mathcal{N}(0, 1)$.

Exemple. *Fonctions de densité $\mathcal{N}_{\mu,\sigma^2}$ des lois normales $\mathcal{N}(0, 1)$ et $\mathcal{N}(2.0, 0.5)$.*



On peut passer d'une loi normale centrée réduite à n'importe quelle loi normale : on multiplie par l'écart-type σ puis on ajoute la moyenne μ.

Proposition 4.27. Soit $\mu \in \mathbb{R}$ et $\sigma > 0$, alors nous avons

$$\begin{array}{ccc} \omega \sim \mathcal{N}(0,1) & \implies & \mu + \sigma \cdot \omega \sim \mathcal{N}(\mu, \sigma^2) \\ \omega \sim \mathcal{N}(\mu, \sigma^2) & \implies & \frac{\omega - \mu}{\sigma} \sim \mathcal{N}(0,1) \end{array}$$

En d'autres mots, depuis n'importe quelle loi normale de moyenne μ et d'écart-type σ, on obtient une loi normale centrée réduite en retranchant la moyenne μ, puis en divisant par l'écart-type σ.
Voici l'espérance et la variance d'une variable aléatoire suivant une loi normale.

Proposition 4.28. Soit $\omega \sim \mathcal{N}(\mu, \sigma^2)$ avec $\mu \in \mathbb{R}$ et $\sigma > 0$. Nous avons

$$\mathop{\mathbb{E}}_{\omega \sim \mathcal{N}(\mu,\sigma^2)} [\omega] = \mu$$

Proposition 4.29. Soit $\omega \sim \mathcal{N}(\mu, \sigma^2)$ avec $\mu \in \mathbb{R}$ et $\sigma > 0$. Nous avons

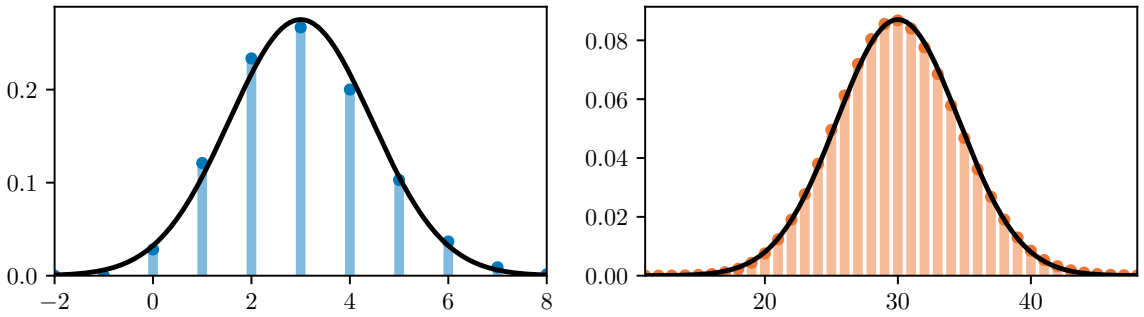
$$\mathop{\mathbb{V}}_{\omega \sim \mathcal{N}(\mu,\sigma^2)} [\omega] = \sigma^2$$

Cette loi est connue car elle nous permet d'*approximer la somme de variables aléatoires i.i.d.* suivant une distribution quelconque μ. En effet, le théorème central limite de Lindeberg–Lévy permet de justifier cette approximation. Voici un énoncé très informel de ce théorème que l'on verra en détail plus tard.

Théorème 4.30 (informel). Soit une séquence $\omega_1, \omega_2, \cdots \sim \mu$ où $\mathbb{V}_{\omega \sim \mu} [\omega] < \infty$. Lorsque $m \to +\infty$ nous avons

$$\sum_{i=1}^{m} \omega_i \sim \mathcal{N}\left(m \mathop{\mathbb{E}}_{\omega \sim \mu} [\omega],\ m \mathop{\mathbb{V}}_{\omega \sim \mu} [\omega] \right).$$

Exemple. *On peut par exemple approximer la somme de variables aléatoires de Bernoulli $\mathcal{B}(p)$ grâce à une loi normale. Pour $p = 0.3$, voici l'approximation lorsque $m = 10$ et $m = 100$.*



Chapitre 5

Simulation de lois

Dans ce chapitre, nous allons étudier comment simuler des lois sur un ordinateur. Nous allons voir plusieurs types d'algorithmes.

- *Algorithmes pour tirer selon $\mathcal{U}(0,1)$:*

  Ils permettent de créer des variables aléatoires sur un ordinateur.

- *Méthodes d'inversion :*

  Ces algorithmes permettent de tirer selon une loi pour laquelle l'inverse de la fonction de répartition est connu.

- *Algorithmes utilisant l'astuce de la reparamétrisation :*

  Ils permettent de tirer selon une loi grâce au tirage d'une autre loi et en reparamétrisant la variable aléatoire.

- *Méthodes du rejet :*

  Ces algorithmes permettent de tirer selon une loi grâce à plusieurs tirages d'une autre loi et en rejetant certains tirages.

5.1 Le cas "simple" : la loi uniforme

Voici deux algorithmes très connus pour simuler la loi uniforme :

- Mersenne Twister
- Blum Blum Shub

Dans la réalité, on va utiliser des fonctions comme `random.random()` dans Python (qui utilise l'algorithme Mersenne Twister).

5.2 Méthodes d'inversion

Fonction de répartition

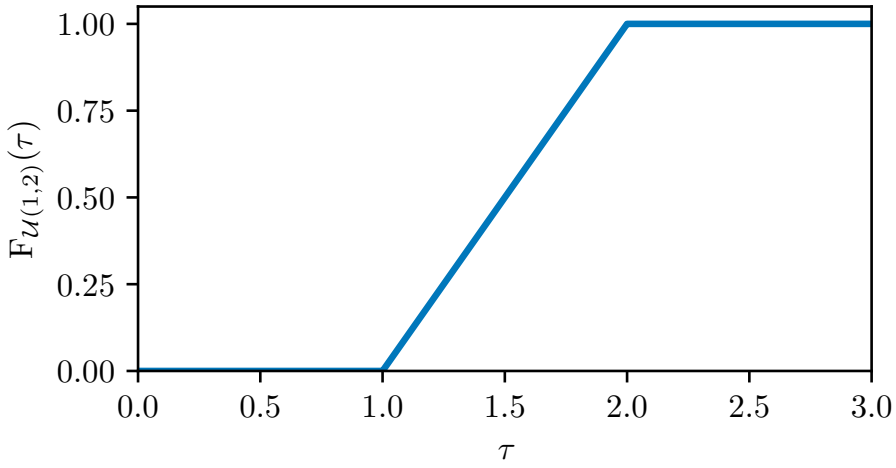
Afin d'évoquer les méthodes d'inversion, il est important de définir les fonctions de répartition (ainsi que de voir leurs propriétés).

Définition 5.1. Soit $\omega \sim \mu$ une variable aléatoire réelle, *i.e.*, μ est une distribution sur $\mathbb{R}$. On appelle fonction de répartition associée à μ la fonction $\mathrm{F}_\mu : \mathbb{R} \to [0,1]$ définie par :

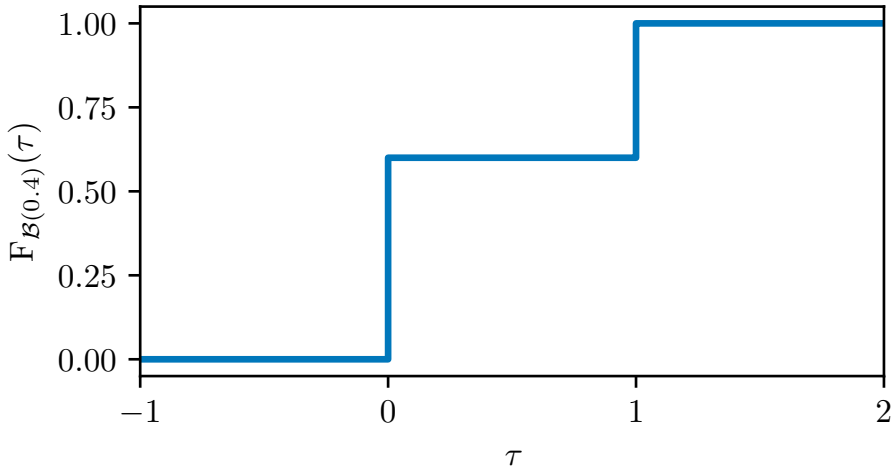
$$\forall \tau \in \mathbb{R}, \quad \mathrm{F}_\mu(\tau) = \mathop{\mathbb{P}}_{\omega \sim \mu} [\omega \leq \tau] .$$

La fonction de répartition F permet de connaître la probabilité que la variable $\omega \sim \mu$ prenne une valeur inférieure ou égale à un seuil donné τ.

Exemple. *Voici la fonction de répartition de la loi* $\mathcal{U}(1,2)$.



Exemple. *Voici la fonction de répartition de la loi* $\mathcal{B}(0.4)$.



La fonction de répartition possède plusieurs propriétés importantes.

Proposition 5.2. Soit μ une distribution sur $\mathbb{R}$ et $\mathrm{F}_\mu : \mathbb{R} \to [0,1]$ sa fonction de répartition. Nous avons

1. F_μ est croissante ;
2. F_μ est continue à droite ;
3. $\lim_{\tau\to-\infty} \mathrm{F}_\mu(\tau) = 0$ et $\lim_{\tau\to+\infty} \mathrm{F}_\mu(\tau) = 1$.

Voici des propriétés qui permettent de montrer les liens forts qui existent entre une distribution et une fonction de répartition.

Proposition 5.3. Soit F une fonction de $\mathbb{R}$ dans $[0,1]$, croissante, continue à droite avec $\lim_{\tau\to-\infty} \mathrm{F}(\tau) = 0$ et $\lim_{\tau\to+\infty} \mathrm{F}(\tau) = 1$. Alors, il existe une mesure de probabilité μ sur $\mathbb{R}$ dont la fonction de répartition est F.

Proposition 5.4. Soient μ_1 et μ_2 des distributions sur $\mathbb{R}$. $\mu_1 = \mu_2$ si et seulement si $F_{\mu_1} = F_{\mu_2}$.

Simulation par inversion

Si F est inversible et que nous avons une expression de son inverse, on peut générer une variable selon la loi F.

Algorithm 1 Simulation par inversion

Tirer $u \sim \mathcal{U}(0,1)$
return $\mathrm{F}_\mu^{-1}(u)$

Exemple. *Considérons une variable aléatoire $\omega \sim \mathcal{E}(\lambda)$ de loi exponentielle de paramètre λ. Sa fonction de répartition est :*

$$F_{\mathcal{E}(\lambda)}(\tau) = 1 - e^{-\lambda\tau}.$$

L'inverse de cette fonction est :

$$F^{-1}_{\mathcal{E}(\lambda)}(u) = -\frac{1}{\lambda}\ln(1-u).$$

Ainsi, nous avons l'algorithme suivant.

Algorithm 2 Simulation d'une loi exponentielle

Tirer $u \sim \mathcal{U}(0,1)$
return $-\frac{\ln(1-u)}{\lambda}$

Remarquons que $1-u \sim \mathcal{U}(0,1)$ lorsque $u \sim \mathcal{U}(0,1)$. On peut donc simplifier l'algorithme, ce qui donne une version "optimisée".

Algorithm 3 Simulation "optimisée" d'une loi exponentielle

Tirer $u \sim \mathcal{U}(0,1)$
return $-\frac{\ln u}{\lambda}$

Fonction inverse généralisée

Au lieu de prendre l'inverse F^{-1} qui n'existe pas forcément, on considère la *fonction inverse généralisée* F^- de F. Il s'agit de l'inverse continue à gauche de F.

Définition 5.5. Soient une distribution μ et sa fonction de répartition $\mathrm{F}_\mu : \mathbb{R} \to [0,1]$. La fonction inverse généralisée est définie par

$$\forall u \in]0,1[, \quad F_\mu^-(u) = \inf\{\tau \in \mathbb{R} \mid F(\tau) \geq u\}, \quad .$$

Voici les différentes propriétés de la fonction inverse généralisée.

Proposition 5.6. Soient une distribution μ avec sa fonction de répartition $\mathrm{F}_\mu : \mathbb{R} \to [0,1]$. Si F_μ est bijective, alors $F_\mu^- = \mathrm{F}_\mu^{-1}$.

Proposition 5.7. Soient une distribution μ avec sa fonction de répartition $\mathrm{F}_\mu : \mathbb{R} \to [0,1]$, $u \in]0,1[$, et $\tau \in \mathbb{R}$, alors $F_\mu^-(u) \leq \tau$ si et seulement si $u \leq \mathrm{F}(\tau)$.

Simulation par la fonction inverse généralisée

On peut maintenant généraliser[1] l'algorithme en utilisant la fonction inverse généralisée.

Algorithm 4 Simulation par la fonction inverse généralisée

Tirer $u \sim \mathcal{U}(0,1)$
return $F_\mu^-(u)$

Nous pouvons ensuite prouver que cette algorithme est correct autrement dit que $F_\mu^-(u) \sim \mu$ si $u \sim \mathcal{U}(0,1)$.

Proposition 5.8. L'algorithme de simulation est correct.

[1] Désolé pour ce jeu de mots...

Démonstration. Soit $\omega' \sim \mu'$ la variable retournée par l'algorithme, c'est-à-dire $\omega' = F_\mu^-(u)$ avec $u \sim \mathcal{U}(0,1)$. Montrons que $\mathrm{F}_{\mu'}(\tau) = \mathrm{F}_\mu(\tau)$ pour tout $\tau \in \mathbb{R}$:

$$\begin{aligned}\mathrm{F}_{\mu'}(\tau) &= \mathop{\mathbb{P}}_{\omega\sim\mu'}\left[\omega' \leq \tau\right] \\ &= \mathop{\mathbb{P}}_{u\sim\mathcal{U}(0,1)}\left[F_\mu^-(u) \leq \tau\right] \\ &= \mathop{\mathbb{P}}_{u\sim\mathcal{U}(0,1)}\left[u \leq F_\mu(\tau)\right] \\ &= \mathrm{F}_\mu(\tau).\end{aligned}$$

□

Voici une liste de distribution avec leur fonction de répartition et leur inverse généralisée qui peuvent servir à l'algorithme pour simuler un tirage.

Distribution μ	$F_\mu(\tau)$	$F_\mu^-(u)$
Uniforme $\mathcal{U}(a,b)$	$\frac{\tau-a}{b-a}$ pour $\tau \in [a,b]$	$a + u(b-a)$
Exponentielle $\mathcal{E}(\lambda)$	$1 - e^{-\lambda\tau}$ pour $\tau \geq 0$	$-\frac{1}{\lambda}\ln(1-u)$
Normale $\mathcal{N}(\mu,\sigma^2)$	$\frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\tau-\mu}{\sigma\sqrt{2}}\right)\right]$	$\mu + \sigma\sqrt{2}\cdot\mathrm{erf}^{-1}(2u-1)$ (Pas calculable)
Poisson $\mathcal{P}(\lambda)$	$e^{-\lambda}\sum_{k=0}^{\tau}\frac{\lambda^k}{k!}$ pour $\tau \in \mathbb{N}$	Pas de forme close
Géométrique $\mathcal{G}(p)$	$1-(1-p)^\tau$ pour $\tau \in \mathbb{N}^*$	$\left\lceil\frac{\ln(1-u)}{\ln(1-p)}\right\rceil$
Binomiale $\mathcal{B}(n,p)$	$\sum_{k=0}^{\tau}\binom{n}{k}p^k(1-p)^{n-k}$ pour $\tau \in \mathbb{N}$	Pas de forme close
Weibull $\mathcal{W}(k,\lambda)$	$1 - e^{-(\tau/\lambda)^k}$ pour $\tau \geq 0$	$\lambda(-\ln(1-u))^{1/k}$
Cauchy $\mathcal{C}(x_0,\gamma)$	$\frac{1}{\pi}\tan^{-1}\left(\frac{\tau-x_0}{\gamma}\right) + \frac{1}{2}$	$x_0 + \gamma\tan[\pi(u-0.5)]$

5.3 Algorithme de Box-Muller : tirer selon $\mathcal{N}(0,1)$

Voici une solution pour générer la loi normale : l'algorithme de Box-Muller. Ce dernier provient d'une interprétation probabiliste du calcul de l'intégrale de Gauss . Cet algorithme est en particulier utilisé dans `std::normal_distribution` de la bibliothèque Standard C++.
On considère deux variables $x_1 \sim \mu_1$ et $x_2 \sim \mu_2$ de loi normale centrée réduite indépendantes $\mathcal{N}(0,1)$, avec les densités suivantes :

$$\begin{aligned}\mu_1(x_1) &= \frac{1}{\sqrt{2\pi}}e^{-\frac{x_1^2}{2}}, \\ \mu_2(x_2) &= \frac{1}{\sqrt{2\pi}}e^{-\frac{x_2^2}{2}}.\end{aligned}$$

Comme $x_1 \sim \mu_1$ et $x_2 \sim \mu_2$ sont indépendantes, la densité conjointe du couple $(x_1,x_2) \sim \mu_1\otimes\mu_2$ est le produit des densités de chaque variable :

$$\begin{aligned}\mu_1\otimes\mu_2(x_1,x_2) &= \mu_1(x_1)\otimes\mu_1(x_2) \\ &= \frac{1}{2\pi}e^{-\frac{x_1^2+x_2^2}{2}}.\end{aligned}$$

L'idée est d'utiliser les coordonnées polaires. Comme (x_1,x_2) est un point aléatoire, on peut parler de ses coordonnées polaires (r,θ), qui sont elles-mêmes des variables aléatoires :

$$\begin{aligned}x_1 &= r\cos(\theta), \\ x_2 &= r\sin(\theta).\end{aligned}$$

Ensuite, en faisant des changements de variables, on peut voir que nous avons

$$\begin{aligned}\int_{\mathbb{R}}\int_{\mathbb{R}} \frac{1}{2\pi} e^{-\frac{x_1^2+x_2^2}{2}} dx_1 dx_2 &= \int_{\mathbb{R}}\int_{\mathbb{R}} \frac{1}{2\pi} e^{-\frac{r^2}{2}} r dr d\theta \\ &= \left[\int_{\mathbb{R}} e^{-\frac{r^2}{2}} r dr\right]\left[\int_{\mathbb{R}} \frac{1}{2\pi} d\theta\right] \\ &= \left[\int_{\mathbb{R}} \frac{1}{2} e^{-\frac{s}{2}} ds\right]\left[\int_{\mathbb{R}} \frac{1}{2\pi} d\theta\right].\end{aligned}$$

Étant donné la densité de (x_1, x_2), on voit que l'angle θ suit une loi uniforme :

$$\theta \sim \mathcal{U}(0, 2\pi).$$

De plus, le rayon $s = r^2$ suit une distribution exponentielle :

$$r^2 \sim \mathcal{E}\left(\frac{1}{2}\right).$$

Transformation de Box-Muller

Ainsi, grâce à ça, nous avons la proposition suivante.

Proposition 5.9. Soient $r^2 \sim \mathcal{E}\left(\frac{1}{2}\right)$ et $\theta \sim \mathcal{U}(0, 2\pi)$. Alors, nous avons

$$r\cos(\theta) \sim \mathcal{N}(0,1) \quad \text{et} \quad r\sin(\theta) \sim \mathcal{N}(0,1)$$

Algorithme de Box-Muller

Grâce à la proposition, nous pouvons en déduire un algorithme pour tirer une variable aléatoire suivant une loi normale.

Algorithm 5 Algorithme de Box-Muller

Tirer $\theta \sim \mathcal{U}(0, 2\pi)$ et $r^2 \sim \mathcal{E}(\frac{1}{2})$
return $r\cos(\theta)$

5.4 Astuce de la reparamétrisation

Algorithme

L'astuce de la reparamétrisation permet de tirer selon une distribution pour laquelle nous n'avons pas d'algorithme pour effectuer un tirage en utilisant une autre distribution. Cette algorithme a été trouvé par (Devroye, 1996).

Algorithm 6 Simulation par l'astuce de la reparamétrisation

Tirer $\omega \sim \mu$
return $f(\omega)$

Exemples

Exemple. *Voici un exemple simple pour tirer selon $\mathcal{U}(0, a)$.*

Algorithm 7 Algorithme pour tirer $\mathcal{U}(0, a)$

Tirer $u \sim \mathcal{U}(0, 1)$
return $a \cdot u$

Exemple. *Voici un autre exemple pour tirer selon $\mathcal{E}(\frac{1}{2})$.*

Algorithm 8 Algorithme pour tirer $\mathcal{E}(\frac{1}{2})$

Tirer $u \sim \mathcal{U}(0, 1)$
return $-2\ln(u)$

On peut en déduire un autre exemple (qui est connu ...).

Exemple. *Nous avons déjà vu un algorithme utilisant cette astuce : l'algorithme de Box-Muller.*

Algorithm 9 Algorithme de Box-Muller

Tirer $u_1 \sim \mathcal{U}(0,1)$ et $u_2 \sim \mathcal{U}(0,1)$
return $\sqrt{-2\ln(u_1)}\cos(2\pi u_2)$

Dans le chapitre précédent, nous avons vu la proposition suivante.

Proposition. Soit $\mu \in \mathbb{R}$ et $\sigma > 0$, alors nous avons

$$\omega \sim \mathcal{N}(0,1) \quad \Longrightarrow \quad \mu + \sigma \cdot \omega \sim \mathcal{N}(\mu, \sigma^2)$$

Elle nous permet de directement obtenir un algorithme utilisant l'astuce de la reparamétrisation.

Exemple. *En effet, nous sommes en mesure de tirer (simplement) une variable aléatoire selon* $\mathcal{N}(\mu, \sigma^2)$.

Algorithm 10 Algorithme pour tirer selon $\mathcal{N}(\mu, \sigma^2)$

Tirer $\omega \sim \mathcal{N}(0,1)$ (avec l'algorithme de Box-Muller)
return $\mu + \sigma \cdot \omega$

Voici plusieurs exemples de fonction f qui peut être utilisée pour faire des tirages. D'ailleurs, on définit la distribution $\mathcal{U}(0,1)^{\otimes 2}$ par $\mathcal{U}(0,1)^{\otimes 2} = \mathcal{U}(0,1) \otimes \mathcal{U}(0,1)$.

Cible	Source	Fonction f
Normal $\mathcal{N}(0,1)$	$(u_1, u_2) \sim \mathcal{U}(0,1)^{\otimes 2}$	$\sqrt{-2\ln(u_1)}\cos(2\pi u_2)$
Normal $\mathcal{N}(\mu, \sigma^2)$	$\omega \sim \mathcal{N}(0,1)$	$\mu + \sigma \cdot \omega$
Exponential $\mathcal{E}(1)$	$u \sim \mathcal{U}(0,1)$	$\ln(1/u)$
Log-Normal $\ln\mathcal{N}(\mu, \sigma^2)$	$\omega \sim \mathcal{N}(\mu, \sigma^2)$	$\exp(\omega)$
Cauchy $\mathcal{C}(0,1)$	$u \sim \mathcal{U}(0,1)$	$\tan(\pi u)$
Rademacher $\mathcal{R}(\frac{1}{2})$	$\omega \sim \mathcal{B}(\frac{1}{2})$	$2\omega - 1$

5.5 Méthode du rejet

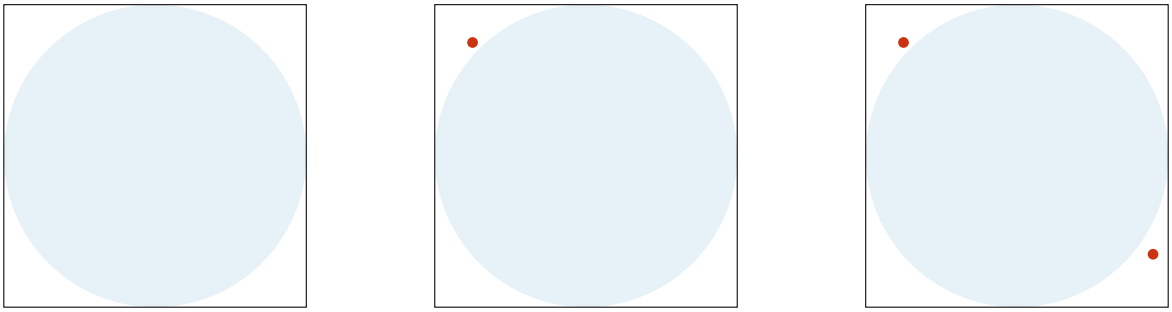
Cas simple : Loi uniforme sur le disque unité

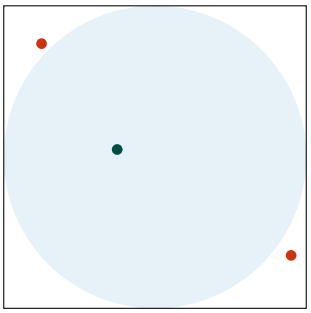
L'objectif est de générer un point uniformément distribué dans le disque unité en utilisant la méthode du rejet.

Algorithm 11 Générer un point dans le disque unité par la méthode du rejet

Générer uniformément un point x dans $[-1,1]^2$
while x n'est pas dans le disque unité **do**
    Générer uniformément un point x dans $[-1,1]^2$
end while
return x

Illustration de la méthode de rejet





Analyse du nombre d'itérations

On peut analyser son nombre d'itérations en moyenne.

Proposition 5.10. Le nombre d'itérations suit la loi géométrique $\mathcal{G}(\frac{\pi}{4})$, et son espérance est donc $\frac{4}{\pi}$.

Démonstration. Un succès correspond à la génération d'un point à l'intérieur du disque unité. La probabilité de succès est donnée par :

$$P(\text{succès}) = \frac{\text{aire du disque}}{\text{aire du carré}} = \frac{\pi}{4}.$$

L'algorithme suit donc un processus où l'on répète les essais jusqu'à obtenir un succès, ce qui correspond à une loi géométrique. □

Théorème 5.11. L'algorithme génère bien une distribution uniforme sur le disque unité.

Démonstration. Cf TD ? □

Méthode du rejet pour une densité à support compact

Problème : Soit une densité $f(x)$ à support compact sur $[a, b]$ et bornée par M. On veut générer une variable de densité f.

Algorithm 12 Méthode du rejet pour une densité bornée

Générer (x, y) uniformément dans $[a, b] \times [0, M]$
while $y > f(x)$ **do**
    Générer (x, y) uniformément dans $[a, b] \times [0, M]$
end whilereturn x

Algorithme du rejet généralisé

Arrêtons maintenant de supposer que f n'est plus à support compact. On suppose maintenant que l'on dispose d'une autre fonction de densité g dont on sait générer des tirages et tel qu'il existe un réel $a \in \mathbb{R}_+$, pour tout $x \in \mathbb{R}$, $f(x) \leq ag(x)$.
Remarquons que $a > 1$ car f et g sont des fonctions de densité (d'intégrale égale à 1). Le cas $a = 1$ signifie que $f = g$, mais ce cas-là est trivial !
Problème :

- une fonction de densité f ;
- une fonction de densité g telle que l'on peut générer des nombres selon la loi de densité g ;
- un nombre $a > 1$ avec $f(x) \leq ag(x)$.

Trouver un générateur pour la distribution de densité f.

Algorithme du rejet généralisé

On peut obtenir l'algorithme suivant.

Algorithm 13 Algorithme du rejet généralisé

Générer x suivant une loi de densité g
Générer un nombre y uniformément dans l'intervalle $[0, ag(x)]$
while $y > f(x)$ **do**
    Générer x suivant une loi de densité g
    Générer un nombre y uniformément dans l'intervalle $[0, ag(x)]$
end while
return x

Partie II

Statistiques

Chapitre 6

Estimation

6.1 Introduction

Hypothèses

À partir de maintenant, la mesure de probabilité / distribution n'est plus considérée comme connue. Nous allons effectuer des estimations afin d'en déduire des informations. On se place dans le cadre paramétré, c'est-à-dire que nous connaissons la forme générale de la loi, mais les paramètres restent inconnus.

Hypothèse 6.1. Nous avons une distribution $\mathcal{D}_{\boldsymbol{\theta}}$, paramétrée par des paramètres $\boldsymbol{\theta} \in \Theta$ inconnus, sur l'ensemble des données X.

Exemple. *Par exemple, on sait que le résultat d'un lancer de pièce suit une loi de Bernoulli, notée $\mathcal{B}(p)$, mais le paramètre p (la probabilité d'obtenir pile) est inconnu.*

Exemple. *Phénomène suivant une loi normale, mais avec des paramètres $\boldsymbol{\theta} = (\mu, \sigma^2)$ (la moyenne et la variance) inconnus.*

Nous considérons également que nous avons un échantillon (c'est-à-dire des réalisations) venant de cette distribution $\mathcal{D}_\theta$ (dont les paramètres sont inconnus).

Définition 6.1 (Échantillon aléatoire / données)**.** Un *échantillon aléatoire* ou des *données* est une suite de variables aléatoires $x_1 \sim \mathcal{D}_{\boldsymbol{\theta}}, \dots, x_m \sim \mathcal{D}_{\boldsymbol{\theta}}$ indépendantes et identiquement distribuées (*i.i.d.*), *i.e.*, nous avons $(x_1, \dots, x_m) \sim \mathcal{D}_{\boldsymbol{\theta}}^{\otimes m}$.

Voici plusieurs exemples d'échantillon.

Exemple (Pièce de monnaie)**.** *Considérons une pièce de monnaie dont la probabilité d'obtenir pile (i.e., d'obtenir un succès ✔) est θ. Pour modéliser les lancers, on considère un échantillon aléatoire tel que*

$$(x_1, \dots, x_m) \sim \mathcal{B}(\theta)^{\otimes m}.$$

Exemple (Longueur d'un animal)**.** *On suppose que la longueur des sardines est modélisée par une loi normale de parametres (inconnues) $\boldsymbol{\theta} = (\mu, \sigma^2)$ où $\mu \in \mathbb{R}$ est la moyenne et $\sigma^2 \in \mathbb{R}^+$ la variance. Ainsi, nous avons*

$$(x_1, \dots, x_m) \sim \mathcal{N}(\boldsymbol{\theta})^{\otimes m} = \mathcal{N}(\mu, \sigma^2)^{\otimes m}.$$

Exemple (Hauteur de personnes)**.** *La hauteur des personnes peut dépendre du sexe. On modélise ce phénomène par une* loi de mélange *de deux lois normales :*

$$(x_1, \dots, x_m) \sim \left(0.5\,\mathcal{N}(\mu_1, \sigma_1^2) + 0.5\,\mathcal{N}(\mu_2, \sigma_2^2)\right) \text{, où } \boldsymbol{\theta} = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2) \text{ sont les paramètres.}$$

6.2 Estimateur

Définition

Un *estimateur* prend les données $(x_1, \dots, x_m) \sim \mathcal{D}_\theta^{\otimes m}$ et fournit une estimation *d'un unique paramètre* $\theta \in \mathbb{R}$.

Définition 6.2 (Estimateur lorsque m est fixe)**.** Quand le nombre m de données est fixé, un *estimateur* de $\theta \in \mathbb{R}$ est une fonction

$$\hat{\theta}_m : \mathbb{R}^m \to \mathbb{R}.$$

Exemple. *Pour estimer l'espérance $\mathbb{E}_{x\sim\mathcal{D}_\theta}[x]$, on peut utiliser l'estimateur*

$$\hat{\theta}_n(x_1,\dots,x_m)=\frac{x_1+\cdots+x_m}{m}.$$

Lorsque la taille de l'échantillon varie, nous pouvons également fournir une estimation d'un paramètre $\theta\in\mathbb{R}$.

Définition 6.3 (Estimateur lorsque m est variable)**.** Lorsque m est variable, un *estimateur* de $\theta\in\mathbb{R}$ est une suite de fonctions

$$\left(\hat{\theta}_m\right)_{m\in\mathbb{N}},\quad \text{où}\quad \hat{\theta}_m:\mathbb{R}^m\to\mathbb{R}.$$

Exemple. *La suite d'estimateurs*

$$\hat{\theta}_m(x_1,\dots,x_m)=\frac{1}{m}\sum_{i=1}^{m}x_i$$

est un estimateur de l'espérance $\mathbb{E}_{x\sim\mathcal{D}_\theta}[x]$.

6.3 Estimateur sans biais

Définition

Afin que l'estimation soit efficace, nous avons besoin que l'estimateur soit sans biais. Voici la définition d'un estimateur sans biais.

Définition 6.4 (Estimateur sans biais)**.** L'estimateur $\hat{\theta}_m$ est *sans biais / non biaisé* si, pour tout $\theta\in\Theta$,

$$\mathop{\mathbb{E}}_{(x_1,\dots,x_m)\sim\mathcal{D}_\theta^{\otimes m}}\left[\hat{\theta}_m(x_1,\dots,X_m)\right]=\theta.$$

Autrement dit, si le paramètre réel vaut θ, l'espérance de l'estimateur $\hat{\theta}_m(x_1,\dots,X_m)$ est égale à l'espérance de la variable aléatoire $x\sim\mathcal{D}_\theta$.

Exemples

Exemple (Moyenne)**.** *Soit $(x_1,\dots,x_m)\sim\mathcal{D}_\theta^{\otimes m}$ un échantillon aléatoire dont la moyenne inconnue est θ. Alors,*

$$\begin{aligned}\mathop{\mathbb{E}}_{(x_1,\dots,x_m)\sim\mathcal{D}_\theta^{\otimes m}}\left[\hat{\theta}_n(x_1,\dots,x_m)\right]&=\mathop{\mathbb{E}}_{(x_1,\dots,x_m)\sim\mathcal{D}_\theta^{\otimes m}}\left[\frac{x_1+\cdots+x_m}{m}\right]\\&=\frac{1}{m}\left(\mathop{\mathbb{E}}_{x_1\sim\mathcal{D}_\theta}[x_1]+\cdots+\mathop{\mathbb{E}}_{x_m\sim\mathcal{D}_\theta}[x_m]\right)\\&=\theta.\end{aligned}$$

Ainsi, l'estimateur

$$(x_1,\dots,x_m)\mapsto\frac{x_1+\cdots+x_n}{n}$$

est sans biais.

Pour illustrer la notion de biais, considérons l'estimation de la variance.

Exemple. *Rappelons que*

$$\mathop{\mathbb{V}}_{x\sim\mathcal{D}_\theta}(x)=\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\theta}\left[\left(x-\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\theta}[x]\right)^2\right].$$

Une estimation naïve de la variance $\theta=\sigma^2$ est

$$\hat{\theta}_m(x_1,\dots,x_m)=\frac{1}{m}\sum_{i=1}^{m}(x_i-\overline{x})^2,\quad \textit{où}\quad \bar{x}=\frac{1}{m}\sum_{i=1}^{m}x_i.$$

Cependant, cet estimateur est biaisé :

$$\mathop{\mathbb{E}}_{(x_1,\dots,x_m)\sim\mathcal{D}_\theta^{\otimes m}}\left[\hat{\theta}_n(x_1,\dots,x_m)\right] = \frac{m-1}{m}\sigma^2.$$

Pour obtenir un estimateur sans biais, on définit :

$$\hat{\theta}'_m(x_1,\dots,x_m) = \frac{m}{m-1}\hat{\theta}_m(x_1,\dots,x_m) = \frac{1}{m-1}\sum_{i=1}^{m}(x_i-\overline{x})^2.$$

6.4 Maximum de vraisemblance

La vraisemblance

La vraisemblance mesure la compatibilité des données $(x_1,\dots,x_m)\sim\mathcal{D}_{\boldsymbol{\theta}}$ avec la loi de paramètres $\boldsymbol{\theta}$. L'idée est de rendre les probabilités d'observer chaque donnée x_i aussi grandes que possible.

Définition 6.5 (Vraisemblance pour une loi discrète ou continue)**.** Soit $\boldsymbol{\theta}$ les paramètres d'une loi discrète ou continue $\mathcal{D}_{\boldsymbol{\theta}}$ (ayant une densité) et $(x_1,\dots,x_m)\sim\mathcal{D}_{\boldsymbol{\theta}}$ des données. La *vraisemblance* est définie par

$$\mathcal{L}_{\mathcal{D}_{\boldsymbol{\theta}}}(x_1,\dots,x_m) = \prod_{i=1}^{m}\mathcal{D}_{\boldsymbol{\theta}}(x_i).$$

Exemple. *Soit $\mathcal{D}_\theta=\mathcal{B}(\theta)$ une loi de Bernoulli de paramètre θ. Pour des données $(x_1,\dots,x_m)\sim\mathcal{B}(\theta)^{\otimes m}$, on a :*

$$\mathcal{L}_{\mathcal{B}(\theta)}(x_1,\dots,x_m) = \theta^{\sum_{i=1}^m x_i}(1-\theta)^{m-\sum_{i=1}^m x_i}.$$

Maximisation de la vraisemblance

Un estimateur des paramètres $\boldsymbol{\theta}$ correspond aux paramètres $\boldsymbol{\theta}$ qui maximise la vraisemblance.

Définition 6.6 (Estimateur du maximum de vraisemblance)**.** Un estimateur du maximum de vraisemblance est une fonction

$$\hat{\boldsymbol{\theta}}_m : \mathbb{R}^m \to \Theta$$

telle que

$$\hat{\boldsymbol{\theta}}_m(x_1,\dots,x_m) \in \mathop{\mathrm{argmax}}_{\boldsymbol{\theta}'\in\Theta}\mathcal{L}_{\mathcal{D}_{\boldsymbol{\theta}'}}(x_1,\dots,x_m).$$

Exemples

Théorème 6.7. Soit $x\sim\mathcal{B}(\theta)$. L'estimateur du maximum de vraisemblance de θ est

$$\hat{\theta}_m(x_1,\dots,x_m) = \frac{1}{m}\sum_{i=1}^{m}x_i.$$

Démonstration. La vraisemblance est

$$\mathcal{L}_{\mathcal{B}(\theta)}(x_1,\dots,x_m) = \theta^{\sum_{i=1}^m x_i}(1-\theta)^{m-\sum_{i=1}^m x_i}.$$

En prenant le logarithme, on obtient

$$\log\mathcal{L}_{\mathcal{B}(\theta)}(x_1,\dots,x_m) = \left(\sum_{i=1}^{m}x_i\right)\log\theta + \left(m-\sum_{i=1}^{m}x_i\right)\log(1-\theta).$$

En dérivant par rapport à θ et en posant la dérivée égale à zéro, on trouve que

$$\theta = \frac{1}{m}\sum_{i=1}^{m}x_i.$$

□

Théorème 6.8. Soit $x \sim \mathcal{N}(\mu, \sigma^2)$. L'estimateur du maximum de vraisemblance de (μ, σ^2) est donné par :

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2.$$

Démonstration. La vraisemblance est

$$\mathcal{L}_{\mathcal{N}(\mu,\sigma^2)}(x_1, \ldots, x_m) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right).$$

En prenant le logarithme et en dérivant par rapport à μ et σ, on obtient respectivement les formules ci-dessus. □

6.5 Méthode des moments

La méthode des moments

La méthode des moments est une alternative pour construire des estimateurs, moins utilisée que le maximum de vraisemblance mais intéressante à connaître.

Définition 6.9. Pour $k \in \mathbb{N}^*$, le *moment d'ordre* k de la variable aléatoire $x \sim \mathcal{D}$ est défini par

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[x^k\right].$$

Exemple. *Si $x \sim \mathcal{E}(\theta)$, alors*

$$\mathop{\mathbb{E}}_{x\sim\mathcal{E}(\theta)}\left[x^k\right] = \frac{k!}{\theta^k}.$$

Estimateur par la méthode des moments

Définition 6.10. Un *estimateur par la méthode des moments* de θ est obtenu en remplaçant un moment théorique d'ordre k par son estimation empirique, c'est-à-dire par

$$\frac{1}{m}\sum_{i=1}^{m} x_i^k.$$

Exemple. *Si $x \sim \mathcal{E}(\theta)$, alors $\mathbb{E}_{x\sim\mathcal{E}(\theta)}[x] = \frac{1}{\theta}$, d'où*

$$\theta = \frac{1}{\mathbb{E}_{x\sim\mathcal{E}(\theta)}[x]}.$$

L'estimateur par la méthode des moments est donc

$$\hat{\theta}_m(x_1, \ldots, x_m) = \frac{m}{\sum_{i=1}^{m} x_i}.$$

Exemple 6.11. *On a également $\mathbb{E}_{x\sim\mathcal{E}(\theta)}[x^2] = \frac{2}{\theta^2}$. On en déduit :*

$$\theta = \sqrt{\frac{2}{\mathbb{E}_{x\sim\mathcal{E}(\theta)}[x^2]}}.$$

Ainsi, un autre estimateur par la méthode des moments est

$$\hat{\theta}_m(x_1, \ldots, x_m) = \sqrt{\frac{2m}{\sum_{i=1}^{m} x_i^2}}.$$

Chapitre 7

Apprentissage statistique

7.1 Introduction

Introduction

Dans le chapitre précédent, nous avons vu le problème de l'estimation en statistique qui est le suivant.

Problème de l'estimation

- **Hypothèses :**
  1. Une distribution $\mathcal{D}_{\boldsymbol{\theta}}$ sur X paramétrée par des paramètres $\boldsymbol{\theta} \in \Theta$ inconnus.
  2. On dispose de données $x_1, \ldots, x_m \sim \mathcal{D}_{\boldsymbol{\theta}}$.
- **Objectif :** Approcher la valeur des paramètres $\boldsymbol{\theta}$ en construisant un estimateur $\hat{\boldsymbol{\theta}} \in \Theta$ qui minimise

$$\min_{\hat{\boldsymbol{\theta}} \in \Theta} \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| \quad \text{où} \quad \|\cdot\| \text{ est une norme sur } \Theta.$$

Nous allons voir comment élargir le problème de l'estimation aux problèmes d'apprentissage statistique que l'on peut rencontrer sur des données où une (simple) estimation ne permet pas de poser un modèle correct sur les données.

Apprentissage en général

L'apprentissage statistique consiste à construire des modèles capables de prédire ou d'expliquer un phénomène à partir des données. Nous pouvons en règle générale, définir l'apprentissage de la manière suivante.[1]

Apprentissage

- **Hypothèses sur les données :**
  1. Une distribution $\mathcal{D}$ sur un ensemble d'exemples Z
  2. Echantillon de données $\mathcal{S} = (z_1, \ldots, z_m) \sim \mathcal{D}^{\otimes m}$
- **Hypothèse sur le modèle :**
  1. Ensemble de modèles $H \subseteq \{h : Z \to Y\}$
- **Hypothèses pour évaluer la qualité du modèle :**
  1. Fonction optimale $h^* : Z \to Y^*$
  2. Fonction de perte $\ell : Y \times Y^* \to \mathbb{R}$ pour évaluer la qualité de $h \in H$ par rapport à h^*
- **Objectif :** Trouver une fonction $h \in H$ afin de minimiser

$$\frac{1}{m} \sum_{i=1}^{m} \ell(h(z_i), h^*(z_i))$$

[1] Attention, c'est indigeste, mais nous allons voir des exemples.

7.2 Apprentissage non supervisé

Apprentissage non supervisé

Nous pouvons spécifier le cadre d'apprentissage à l'apprentissage non supervisé.

Apprentissage non supervisé

- **Hypothèse supplémentaire :** la fonction optimale $h^* : Z \to Y^*$ est soit
  1. inconnue
  2. construite à la main
- **Objectif :** Trouver une fonction $h \in H$ afin de minimiser

$$\frac{1}{m} \sum_{i=1}^{m} \ell(h(z_i), h^*(z_i))$$

Estimation

L'estimation en statistique est un exemple (très simple) d'apprentissage non supervisé.

Problème de l'estimation

- **Hypothèse sur le modèle :**
  1. Ensemble de modèle $H = \{z \mapsto \hat{\boldsymbol{\theta}}(z_1, \ldots, z_m) \mid (z_1, \ldots, z_m) \in Z^m\}$
- **Hypothèses pour évaluer la qualité du modèle :**
  1. Fonction optimale $h^* : z \mapsto \boldsymbol{\theta}$
  2. Fonction de perte $\|\cdot\|$ est la norme sur Θ

Exemples d'apprentissage non supervisé

Voici des exemples d'apprentissage non supervisé (parmi beaucoup d'autres) :

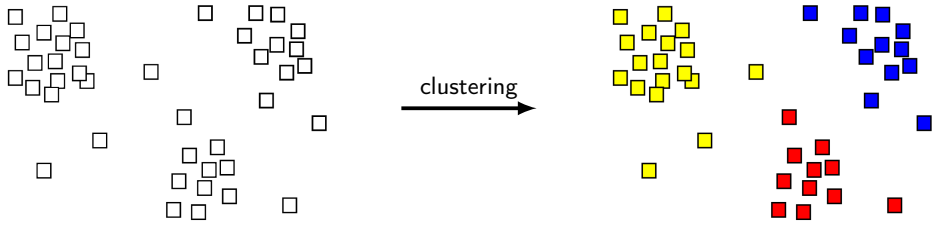
- **Estimation :** Maximisation de la vraisemblance, Méthode des moments ;
- **Clustering :** *Algorithme des k-moyennes*, Algorithme espérance-maximisation ;
- **Visualisation de données :** Analyse en composantes principales, Algorithme t-SNE ;
- **Estimation de densités :** Méthode de Parzen-Rosenblatt.

Nous allons voir un peu plus en détails l'algorithme des k-moyennes.

Problème de clustering

Problème de clustering

Entrée : des données $z_1, \ldots, z_m \sim \mathcal{D}$ où $\mathcal{D}$ est une distribution sur Z
Sortie : une fonction $h : Z \to \{1, \ldots, K\}$ permettant de partitionner les données en K clusters



Définition 7.1 (Partition en clusters). On partitionne l'ensemble des données :

$$\{x_1, \dots, x_m\} = \bigsqcup_{k=1}^{K} C_k,$$

avec $C_k \neq \emptyset$.

Définition 7.2 (Cluster). Chaque ensemble C_k est appelé un cluster.

Définition 7.3 (Centroïde). Le centroïde du cluster C_k est défini par

$$\mu_{C_k} := \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i.$$

Problème de clustering

- **Hypothèse sur le modèle :**
  1. Ensemble de modèle $H \subseteq \cup_{k \in \mathbb{N}^*} \{h : Z \to \{1, \dots, K\}\}$
- **Hypothèses pour évaluer la qualité du modèle :**
  1. Clustering optimal (et inconnu) $h^* : Z \to \{1, \dots, K\}$
  2. Fonction de perte $y, y^* \mapsto \mathbb{1}[y \neq y^*]$ appelé perte 0-1

Algorithme des k-moyennes

Problème des k-moyennes
Entrée : des données $x_1, \dots, x_m \in \mathbb{R}^p$
Sortie : une partition en K clusters $C_1, \dots, C_K$ qui minimise

$$\sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_{C_k}\|_2^2.$$

Algorithme de Lloyd (Pseudo-code)

Algorithme de Lloyd

1. Choisir aléatoirement K centres $\mu_1, \dots, \mu_K$ (parmi les données).
2. **Tant que la partition change** :
   (a) Pour chaque donnée x_i, assigner x_i au cluster C_k tel que $k = \arg\min_j \|x_i - \mu_j\|_2$.
   (b) Pour chaque cluster C_k, mettre à jour le centroïde :

$$\mu_k := \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i.$$

3. Sortie : la partition $\{C_1, \dots, C_K\}$.

7.3 Apprentissage supervisé

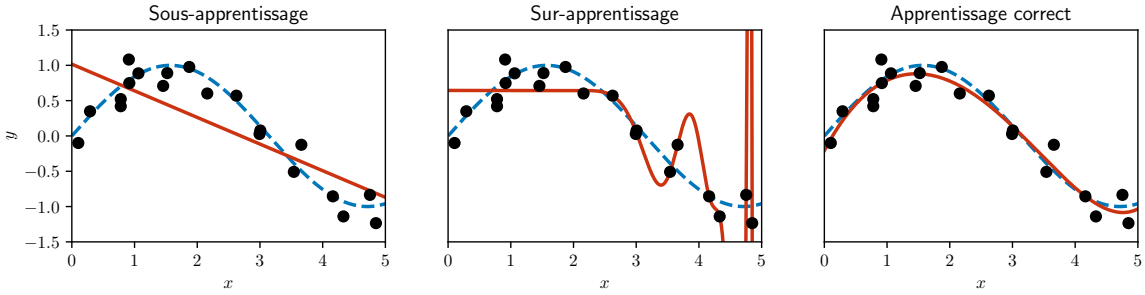
Apprentissage supervisé

Apprentissage supervisé

- **Hypothèses sur les données :**
  1. Un exemple $z_i = (\mathbf{x}_i, y_i) \in X \times Y = Z$ est un couple entrée/sortie
  2. Echantillon de données $\mathcal{S} = (z_1, \dots, z_m) \sim \mathcal{D}^{\otimes m}$
- **Hypothèse sur le modèle :**
  1. Ensemble de modèle $H \subseteq \{h : X \to Y'\}$
- **Hypothèses pour évaluer la qualité du modèle :**
  1. Fonction optimale $(x, y) \mapsto y$
  2. Fonction de perte $\ell : Y' \times Y \to \mathbb{R}$ pour évaluer la qualité de $h \in H$
- **Objectif :** Trouver une fonction $h \in H$ afin de minimiser

$$\frac{1}{m} \sum_{i=1}^{m} \ell(h(\mathbf{x}_i), y_i)$$

Nous ne verrons pas le problème dans le cours, mais lorsque nous apprenons un modèle $h \in H$, nous pouvons rencontrer deux types de problèmes : le sous-apprentissage ou le sur-apprentissage.



Le sous-apprentissage se produit lorsque le modèle est trop simple pour saisir la complexité réelle des données, ce qui entraîne une performance médiocre sur les données. A l'inverse, le sur-apprentissage survient lorsqu'un modèle trop "complexe" s'ajuste excessivement aux données, au point de perdre sa capacité à généraliser sur le problème réel (la distribution $\mathcal{D}$).

Apprentissage supervisé

Il existe beaaaaaauuuuuucoup de modèles ! Voici des exemples séparés en deux types.

- **Modèles paramétriques :** Modèle entièrement décrit par un nombre fixe de paramètres
  - *Régression linéaire* ;
  - Régression logistique ;
  - Réseaux de neuronnes
- **Modèles non-paramétriques :** Modèle non décrit pas des paramètres
  - k-plus proches voisins ;
  - Arbres de décision ;
  - Méthodes à noyaux

Nous allons voir un exemple dans la suite : la régression linéaire.

Régression linéaire

Voici comment nous pouvons présenter le problème de la régression linéaire.

Régression linéaire
Entrée : des données $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m) \in \mathbb{R}^p \times \mathbb{R}$
Sortie : trouver $\boldsymbol{\theta} = (\theta^0, \theta^1, \dots, \theta^p) \in \mathbb{R}^{p+1}$ qui minimise

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} \left\{ \mathcal{J}(\boldsymbol{\theta}) = \sum_{i=1}^{m} \left[y_i - (\theta^0 + \theta^1 x_i^1 + \cdots + \theta^p x_i^p) \right]^2 \right\}.$$

Autrement dit, on cherche à modéliser $y \in \mathbb{R}$ comme une combinaison linéaire des composantes de $\mathbf{x} \in \mathbb{R}^p$.
Nous pouvons trouver la solution du problème d'optimisation de façon analytique. Pour cela, nous avons besoin de représenter les données sous la forme d'une matrice et d'un vecteur.

Définition 7.4 (Matrice des données)**.** La matrice des données s'écrit

$$\mathbf{X} = \begin{pmatrix} 1 & x_1^1 & \dots & x_1^p \\ \vdots & \vdots & & \vdots \\ 1 & x_m^1 & \dots & x_m^p \end{pmatrix} \in \mathbb{R}^{m \times (p+1)} \quad \text{et} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} \in \mathbb{R}^m.$$

Chaque ligne correspond à une observation (avec la première colonne composée de 1 pour le terme constant).
Ensuite, nous pouvons trouver le paramètre optimal pour ces données.

Théorème 7.5. Si $\mathbf{X}^\top \mathbf{X}$ est inversible, alors l'estimateur par moindres carrés est

$$\boldsymbol{\theta}^* = \operatorname*{argmax}_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} \mathcal{J}(\boldsymbol{\theta}) \quad \text{où} \quad \boldsymbol{\theta}^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Démonstration. La fonction à minimiser est

$$\mathcal{J}(\boldsymbol{\theta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}).$$

Le gradient est

$$\nabla_\theta \mathcal{J}(\boldsymbol{\theta}) = -2\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}).$$

En annulant le gradient, on obtient $\mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top \mathbf{X}\boldsymbol{\theta}$, d'où le résultat. □

Remarque 7.6. *Si $\mathbf{X}^\top \mathbf{X}$ n'est pas inversible, on peut utiliser le pseudo-inverse (par exemple celui de Moore-Penrose) ou éliminer les features redondantes.*

Cet estimateur a d'ailleurs une propriété intéressante.

Définition 7.7 (BLUE)**.** Un estimateur linéaire non biaisé $\boldsymbol{\theta}^*$ est dit *BLUE* (Best Linear Unbiased Estimator) s'il possède la plus petite variance parmi tous les estimateurs linéaires non biaisés.

Théorème 7.8 (Théorème de Gauss-Markov)**.** Si nous avons la relation suivante :

$$\forall x \in X, \quad y = \mathbf{x}^T \boldsymbol{\theta} + \varepsilon.$$

avec $\mathbb{E}_{\varepsilon \sim \mu}[\varepsilon] = 0$ et $\mathbb{V}_{\varepsilon \sim \mu}(\varepsilon) = \sigma^2$, l'estimateur

$$\boldsymbol{\theta}^* = \operatorname*{argmax}_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} \mathcal{J}(\boldsymbol{\theta}) \quad \text{où} \quad \boldsymbol{\theta}^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

est le BLUE.

Références

Luc Devroye. Random variate generation in one line of code. *Conference on Winter simulation*. (1996).

Christopher M. Bishop et Nasser M. Nasrabadi. Pattern recognition and machine learning. *Springer*. (2006).

Trevor Hastie, Robert Tibshirani, Jerome H Friedman et Jerome H Friedman. The elements of statistical learning : data mining, inference, and prediction. *Springer*. (2009).

Mehryar Mohri, Afshin Rostamizadeh et Ameet Talwalkar. Foundations of Machine Learning. *MIT Press*. (2012).

Peter Whittle. Probability via expectation. *Springer Science & Business Media*. (2012).

Vladimir Vapnik. The nature of statistical learning theory. *Springer science & business media*. (2013).

Olivier Garet et Aline Kurtzmann. De l'intégration aux probabilités-2e édition augmentée. *Editions Ellipses*. (2019).

Glenn Shafer et Vladimir Vovk. Game-theoretic foundations for probability and finance. *John Wiley & Sons*. (2019).

Philippe Barbe et Michel Ledoux. Probabilité. *Probabilité. EDP Sciences*. (2021).

Kevin P. Murphy. Probabilistic machine learning : an introduction. *MIT press*. (2022).

Francis Bach. Learning theory from first principles. *MIT press*. (2024).