Probabilités et Statistique

Tronc Commun Scientifique Première Année

Florian SIMATOS

2017–2018

Version du 21 novembre 2017.

Remerciements

L'enseignement des probabilités et de la statistique à l'ISAE est une œuvre collective et j'ai eu la chance d'interagir pour élaborer et dispenser ce cours avec de nombreuses personnes. Je tiens à remercier très chaleureusement tous les intervenants, passés et présents, pour leur implication qui fait que ce cours est, année après année, apprécié des étudiants. Je voudrais remercier plus particulièrement Claudie Chabriac pour son enthousiasme indéfectible et communiquant qu'elle partage généreusement avec tous les gens qui la côtoient. Enfin, une pensée amicale pour Patrice Henry qui, pour la première fois depuis de très nombreuses années, ne prendra malheureusement pas part à l'enseignement de cette matière cette année. Patrice, le lien que tu arrives à faire entre cette matière apparemment très théorique et les applications pratiques que tu as rencontrées pendant toute ta carrière au CNES ainsi que la passion que tu arrives à transmettre aux étudiants nous manqueront.

Organisation du cours

Présentation générale

Le cours de probabilités et statistique est divisé en 8 séances de 3 heures et 2 BE de 2 heures, qui se concluent par un examen sur table final de 2 heures. Chaque séance de 3 heures (sauf celle sur les tests d'hypothèse) est divisée en 1 heure de cours magistral (amphi) suivie de 2 heures de PC. La séance sur les tests d'hypothèse consiste quant à elle en 3h de Cours/PC. Les séances 1 à 4 sont consacrées aux probabilités :

Séance 1 : Variables aléatoires discrètes – Chapitre 1 ;

Séance 2 : Variables aléatoires absolument continues – Chapitre 2 ;

Séance 3 : Théorèmes limites – Chapitre 3 ;

Séance 4 : Vecteurs gaussiens – Chapitre 4 ;

les séances 5, 6 et 7 à la statistique inférentielle :

Séances 5 & 6 : Estimation paramétrique – Chapitre 5 ;

Séance 7 : Tests d'hypothèses – Chapitre 6 ;

et la dernière séance est une séance récapitulative :

Séance 8 : Séance panoramique.

Comme indiqué ci-dessus avec la référence aux chapitres correspondant, l'organisation de ce polycopié suit fidèlement ce plan. Par ailleurs, chaque chapitre se conclue par une fiche de synthèse qui regroupe les notions essentielles du chapitre, ainsi que par une liste d'exercices et de problèmes. Les exercices d'application directe du cours sont notés avec une flèche ↪.

Pré-requis

Il est supposé que les étudiants ont déjà une connaissance en probabilités correspondant au programme des classes préparatoires, et sont donc familiers avec la théorie des probabilités discrètes et notamment les notions d'espace de probabilités, variable aléatoire, espérance, variance et indépendance. De manière plus précise, en préalable de ce cours chaque étudiant est supposé connaître le contenu des Sections 1.1 à 1.4 du Chapitre 1 : si tel n'était pas le cas, il est attendu que chaque étudiant se mette au niveau.

Séances 1 à 4 et BE 1 : Probabilités

La première séance de cours consiste en compléter la théorie des probabilités discrètes en introduisant la notion fondamentale d'**espérance conditionnelle** (Sections 1.6 et 1.7)

et en discutant rapidement plusieurs transformées classiques qui ont en fait essentiellement déjà été vues dans le cours de mathématiques déterministes (Section 1.5).

La deuxième séance est consacrée à la théorie des **variables aléatoires absolument continues** : par exemple, cela permettra de parler d'une variable aléatoire uniformément répartie sur l'intervalle $[0, 1]$, qui représente le résultat de l'expérience "Tirer un nombre au hasard dans $[0, 1]$". Le deuxième chapitre du polycopié revisite dans ce cadre toutes les notions introduites dans le cadre discret.

La troisième séance introduit des résultats de convergence de suites de variables aléatoires, qui font le lien entre les probabilités discrètes et absolument continues. On verra par exemple dans quel sens une suite de variables aléatoires discrètes uniformément distribuées sur l'ensemble fini $\{0, 1/n, 2/n, \ldots, 1 - 1/n, 1\}$ converge vers une variable aléatoire absolument continue uniformément répartie sur l'intervalle $[0, 1]$. On verra dans ce chapitre les deux résultats les plus importants de la théorie des probabilités : la **loi forte des grands nombres** et le **théorème central limite**.

La dernière séance est consacrée à l'étude des **vecteurs gaussiens**, qui jouent un rôle très important dans de nombreux domaines de l'ingénierie et notamment le traitement du signal.

Le BE de probabilités vise à illustrer les résultats théoriques vus en cours par des simulations numériques portant sur l'urne de Pólya.

Séances 5 à 7 et BE 2 : Statistique

Pour la plupart des gens, probabilités et statistique sont des notions interchangeables : il s'agit néanmoins bien de domaines différents. De manière simplificatrice, la différence fondamentale est que la statistique est intrinsèquement liée aux données. Cela ne veut pas dire que la statistique est une science purement appliquée : il s'agit en fait d'un domaine très large, qui va de la récolte à l'analyse de données en passant par la conception des expériences, couvrant ainsi des aspects à la fois théoriques et pratiques.

On distingue deux grands domaines de la statistique : la **statistique inférentielle**, et la **statistique descriptive** ou **exploratoire**. La statistique inférentielle postule l'existence d'un modèle probabiliste expliquant les données observées, et vise à remonter aux paramètres de ce modèle à l'aide des observations. Ainsi, il s'agit d'une démarche inverse à celle des probabilités : au lieu de partir d'un modèle probabiliste et d'étudier le comportement théorique des variables aléatoires en jeu, en statistique on part des observations et on essaie de remonter au modèle probabiliste. La statistique inférentielle est donc intrinsèquement liée à la théorie des probabilités, mais le changement de point de vue est tel qu'il correspond à un domaine scientifique très différent.

En statistique descriptive par contre, la théorie des probabilités ne joue pas de rôle : on cherche à *synthétiser, résumer, structurer l'information contenue dans les données dans le but de mettre en évidence des propriétés de l'échantillon et de suggérer des hypothèses* [1].

Comme nous le verrons, les problèmes théoriques de statistique sont en fait des problèmes de probabilités et font appel aux outils probabilistes les plus avancés. Nous verrons notamment dans ce cours que les résultats de convergence – et notamment la loi des grands nombres et le théorème central limite – et les vecteurs gaussiens jouent un rôle importants.

1. Description tirée du livre *Probabilités, Analyse des Données et Statistique* de Gilbert Saporta, auquel le lecteur intéressé pourra se reporter pour une discussion plus détaillée.

Le BE de statistique porte sur le thème des tests d'hypothèse et prend comme exemple d'application la génération de nombres aléatoires.

Séance 8 : Séance panoramique

La dernière séance de cours est une séance récapitulative, où l'on met ensemble toutes les notions abordées. La séance d'exercice est notamment l'occasion d'aborder un problème global mêlant probabilités et statistique.

Table des matières

Notations

Dans tout ce polycopié les notations suivantes seront utilisées : $\mathbb{R}$ désigne l'ensemble des réels, $\mathbb{R}_+ = [0, \infty[$ l'ensemble des réels positifs, $|\Omega| \in \mathbb{N} \cup \{\infty\}$ la cardinalité d'un ensemble (avec $|\Omega| = \infty$ si et seulement si Ω est infini) et $x^+ = \max(x, 0)$ et $x^- = -\min(x, 0)$ les parties positive et négative d'un réel $x \in \mathbb{R}$. On notera $\mathbb{Z}$ l'ensemble des entiers relatifs, $\mathbb{N} = \mathbb{Z} \cap \mathbb{R}_+ = \{0, 1, 2, \ldots\}$ l'ensemble des entiers naturels positifs et $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$ l'ensemble des entiers naturels privé de 0.

Un élément $M \in \mathbb{R}^{m \times n}$ pour $m, n \in \mathbb{N}^*$ est assimilé à une matrice à coefficients dans $\mathbb{R}$ avec m lignes et n colonnes et représente l'application linéaire $x \in \mathbb{R}^n \mapsto Mx \in \mathbb{R}^m$; $M^T \in \mathbb{R}^{n \times m}$ désigne sa transposée. On assimilera $\mathbb{R}^n$ et $\mathbb{R}^{n \times 1}$, i.e., un vecteur $x \in \mathbb{R}^n$ sera vu comme un vecteur colonne avec n lignes et une colonne, et l'on notera donc parfois $x = (x_1, \ldots, x_n) = (x_1 \ \cdots \ x_n)^T \in \mathbb{R}^n$. Pour $x, y \in \mathbb{R}^n$ on notera $\langle x, y \rangle = x^T y = y^T x$ le produit scalaire euclidien entre x et y : $\langle x, y \rangle = \sum_{k=1}^n x_k y_k$ si $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots, y_n)$, et toute inégalité vectorielle $x \leq y$ est à comprendre coordonnée par coordonnée. Pour M une matrice et $a \in \mathbb{R}$, aM désigne la matrice de même taille que M où chaque coordonnée est multipliée par a et rang(M) est son rang. Enfin, pour A une matrice carrée à coefficients dans $\mathbb{R}$, on notera det(A) son déterminant.

Première partie

Théorie des probabilités

Chapitre 1

Variables aléatoires discrètes

On rappelle dans ce chapitre les éléments de la théorie des probabilités discrètes vues en classes préparatoires, que l'on complètera par la notion essentielle d'espérance conditionnelle dans les Sections 1.6 et 1.7.

1.1 Espace de probabilités dans le cas discret

Dans tout ce chapitre, on considère

$$\boxed{\Omega \text{ \textbf{un ensemble dénombrable.}}}$$

Ainsi, Ω est fini ou infini dénombrable. L'ensemble Ω est appelé **univers**, **espace d'états** ou **espace des réalisations** : il est à envisager comme l'ensemble des résultats possibles d'une expérience aléatoire. Puisque l'on se restreint au cas où Ω est dénombrable, on ne considère donc que des expériences "simples", par exemple un lancer de dé. Un élément $\omega \in \Omega$, assimilé au singleton $\{\omega\}$, sera appelé **évènement élémentaire**.

Dans le cadre de ce chapitre où Ω est dénombrable, on considère $\mathcal{F} = \mathcal{P}(\Omega)$ l'ensemble des parties de Ω : comme nous le verrons au Chapitre 2 cela ne sera plus nécessairement le cas lorsque Ω n'est pas dénombrable. L'ensemble $\mathcal{F}$ définit les ensembles que l'on peut mesurer : ainsi, on appellera un ensemble $A \in \mathcal{F}$ **ensemble mesurable** ou **évènement**. Le couple $(\Omega, \mathcal{F})$ est quant à lui appelé **espace mesurable**.

Afin de rendre ces définitions et les suivantes plus concrètes, il sera utile d'avoir un ou plusieurs exemples en tête. Pour cela, examinons comment modéliser plusieurs expériences aléatoires.

Exemple 1.1.1. Un joueur lance deux dés discernables (par exemple, ils ont des couleurs différentes) à 6 faces. La manière la plus simple de modéliser cette expérience aléatoire est donc de considérer

$$\Omega_1 = \{1, \dots, 6\} \times \{1, \dots, 6\}$$

où pour $\omega = (i, j) \in \Omega$, la première coordonnée i donne le résultat du premier dé et la deuxième coordonnée j le résultat du deuxième dé.

Exemple 1.1.2. Considérons maintenant la même expérience, mais dans le cas où les deux dés sont indiscernables. On peut alors enregistrer le résultat de l'expérience de plusieurs manières. On ne peut plus parler de premier et de deuxième dé, mais on peut par exemple enregistrer le plus petit et le plus grand résultat, ce qui nous amène à considérer

$$\Omega_2 = \{(i, j) : 1 \leq i \leq j \leq 6\}.$$

Une autre manière de rendre compte de cette expérience est d'enregistrer le plus petit résultat et la différence entre le plus grand et le plus petit résultat, ce qui nous amène alors à considérer

$$\Omega_3 = \{(i,j) : 1 \leq i \leq 6, 0 \leq j \leq 6-i\}.$$

Exemple 1.1.3. Un autre exemple est donné par un joueur qui joue 100 fois à pile ou face. On peut alors enregistrer le résultat de l'expérience par une suite de longueur 100 de 0/1, où 0 correspond à pile et 1 à face, ce qui correspond à l'univers

$$\Omega_4 = \{0,1\}^{100}.$$

On peut aussi décrire le résultat en enregistrant le résultat du premier lancer (0 on commence par pile, 1 par face) puis la longueur des séries de pile ou face successifs. Par exemple, $(0,3,4,2,91,0,\ldots,0)$ signifie que l'on a d'abord 3 fois pile, puis 4 fois face, puis 2 fois pile et enfin 91 fois face. Ainsi, l'univers correspondant est

$$\Omega_5 = \{0,1\} \times \{0,1,\ldots,100\}^{100}.$$

Les exemples ci-dessus correspondent à des univers finis, même si on pourrait par exemple rendre compte de la première expérience aléatoire par l'univers $\mathbb{N} \times \mathbb{N}$ puisque l'on peut toujours décrire une expérience aléatoire par un univers plus gros. Dans les exemples suivants par contre, on ne peut pas décrire l'expérience aléatoire par un univers fini.

Exemple 1.1.4. Un joueur lance une pièce, et l'on s'intéresse au nombre de lancers nécessaires avant d'obtenir pile pour la première fois : l'univers Ω décrivant cette expérience est alors

$$\Omega_6 = \mathbb{N}^* \cup \{\infty\},$$

l'ajout de l'infini étant là au cas où la pièce soit truquée et ait deux côtés face.

Exemple 1.1.5. Dans le jeu de rôle Earthdawn, le résultat du jet d'un dé à n faces est donné par la règle suivante : tant que l'on obtient le score maximum n, on relance le dé et on additionne le résultat de tous les jets. Ainsi, on peut potentiellement obtenir n'importe quel résultat et un univers naturel est donc

$$\Omega_7 = \mathbb{N}^*.$$

Une autre description de cette expérience aléatoire est d'enregistrer le nombre de fois où l'on obtient le score maximal ainsi que le résultat du dernier lancer : l'univers est alors

$$\Omega_8 = \mathbb{N} \times \{1,\ldots,n-1\}.$$

On retiendra notamment des exemples ci-dessus que

Une même expérience aléatoire peut être décrite par plusieurs univers Ω.

Dans les exemples ci-dessus, la description de l'expérience aléatoire n'est pas complète : par exemple, le résultat d'un lancer de dés sera très différent si l'on jette des dés biaisés ou non. De manière générale, il n'est pas suffisant de décrire l'ensemble des résultats possibles et les évènements – via l'espace mesurable $(\Omega, \mathcal{F})$ – mais il faut aussi décrire les probabilités d'occurrence des divers évènements.

Définition 1.1.1. Une fonction $\mathbb{P} : \mathcal{F} \to [0,1]$ est appelée **mesure de probabilité** si elle vérifie les deux propriétés suivantes :

1. $\mathbb{P}(\Omega) = 1$;
2. pour toute famille dénombrable $(A_i, i \in I)$ d'ensembles mesurables deux à deux disjoints, i.e., I est dénombrable, $A_i \in \mathcal{F}$ pour tout $i \in I$ et $A_i \cap A_j = \emptyset$ pour tout $i \neq j$ dans I, alors
$$\mathbb{P}\left(\bigcup_{i\in I} A_i\right) = \sum_{i\in I} \mathbb{P}(A_i). \tag{1.1}$$

Si $\mathbb{P}$ est une mesure de probabilité sur l'espace mesurable $(\Omega, \mathcal{F})$, le triplet $(\Omega, \mathcal{F}, \mathbb{P})$ est appelé **espace de probabilité**.

Dans le cadre particulier de ce chapitre où Ω est dénombrable, on dira plus spécifiquement que $\mathbb{P}$ est une **mesure de probabilité discrète**, cf. la Section 2.4 pour la définition générale. Le triplet $(\Omega, \mathcal{F}, \mathbb{P})$ décrit maintenant bien une unique expérience aléatoire, et en outre :

Des mesures de probabilité différentes représentent des caractéristiques statistiques différentes !

Pour illustrer ce point revenons à l'exemple 1.1.1 du lancer de 2 dés.

Exemple 1.1.6. Imaginons que les dés sont non biaisés : alors tous les résultats sont équiprobables et puisqu'il y a $6 \times 6 = 36$ résultats possibles, la "bonne" mesure de probabilité $\mathbb{P}_1$ qui décrit cette expérience aléatoire est donc donnée par
$$\mathbb{P}_1(\{\omega\}) = \frac{1}{36}, \ \omega \in \Omega_1.$$
Si en revanche les dés sont biaisés, par exemple chaque dé donne i avec probabilité q_i possiblement $\neq 1/6$, alors la "bonne" mesure de probabilité qui rend compte de cette expérience est donnée par
$$\mathbb{P}'_1(\{(i,j)\}) = q_i q_j, \ (i,j) \in \Omega_1.$$
Ainsi, le couple $(\Omega, \mathcal{F})$ représente l'expérience aléatoire "jeter deux dés à 6 faces", et la mesure de probabilité vient **décrire les caractéristiques statistiques** des dés, ici, $\mathbb{P}_1$ pour des dés non biaisés et $\mathbb{P}_2$ pour des dés possiblement biaisés.

Lorsque l'univers Ω est fini, la probabilité qui à chaque évènement élémentaire associe la probabilité $1/|\Omega|$ est appelée **mesure uniforme** : tous les résultats de l'expérience aléatoire sont équiprobables. Dans ce cas, la probabilité d'un évènement A est proportionnelle à sa cardinalité :
$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}, \ A \in \mathcal{F}.$$

Dans l'exemple ci-dessus, nous avons ramené la description d'une mesure de probabilité discrète à la spécification de la probabilité des évènements élémentaires. Cette approche est correcte car il y a une bijection entre les mesures de probabilité discrètes sur Ω et les suites $(p_\omega, \omega \in \Omega)$ indexées par les évènements élémentaires et satisfaisant
$$\sum_{\omega\in\Omega} p_\omega = 1 \quad \text{et} \quad \forall \omega \in \Omega, p_\omega \geq 0. \tag{1.2}$$

En effet, étant donné $\mathbb{P}$ mesure de probabilité sur $(\Omega, \mathcal{F})$, la suite $p_\omega = \mathbb{P}(\{\omega\})$ satisfait (1.2) puisque $\mathbb{P}$ est à valeurs dans $[0,1]$ et que

$$\begin{aligned}
\sum_{\omega\in\Omega} p_\omega &= \sum_{\omega\in\Omega} \mathbb{P}(\{\omega\}) && \text{(par définition de } p_\omega\text{)} \\
&= \mathbb{P}\left(\bigcup_{\omega\in\Omega}\{\omega\}\right) && \text{(par (1.1) avec } I=\Omega \text{ et } A_\omega=\{\omega\}\text{)} \\
&= \mathbb{P}(\Omega) && \text{(puisque } \Omega = \cup_{\omega\in\Omega}\{\omega\}\text{)} \\
&= 1 && \text{(par définition d'une mesure de probabilité).}
\end{aligned}$$

Réciproquement, étant donnée une suite $(p_\omega, \omega \in \Omega)$ satisfaisant (1.2), on peut définir $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ de la manière suivante :

$$\mathbb{P}(A) = \sum_{\omega\in A} p_\omega, \ \ A \in \mathcal{F}.$$

On vérifie alors aisément que $\mathbb{P}$ est bien une mesure de probabilité, qui satisfait par ailleurs $\mathbb{P}(\{\omega\}) = p_\omega$. On retiendra donc que

Pour décrire une mesure de probabilité discrète $\mathbb{P}$, il suffit de spécifier la famille $(\mathbb{P}(\{\omega\}), \omega \in \Omega)$ des probabilités des évènements élémentaires.

Avant de conclure cette section avec certaines propriétés élémentaires satisfaites par toute mesure de probabilité dans la Proposition 1.1.1 ci-dessous, nous noterons que, pour une même expérience aléatoire, le choix de l'univers Ω influence la "bonne" mesure de probabilité, i.e., celle qui décrit l'expérience aléatoire. Pour comprendre cela revenons à nouveau à l'exemple du lancer de 2 dés.

Exemple 1.1.7. Considérons l'expérience aléatoire consistant à jeter 2 dés discernables et non biaisés. Comme nous l'avons vu dans l'exemple 1.1.6, on peut décrire cette expérience par l'espace de probabilité $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ avec $\mathbb{P}_1$ la mesure uniforme sur Ω_1.
Quitte à perdre l'identité des deux dés, on peut aussi décrire cette expérience en considérant l'univers Ω_2, ce qui revient à enregistrer le plus petit et le plus grand résultat. Dans ce cas, il est clair que la mesure uniforme sur Ω_2 ne rend pas bien compte de l'expérience aléatoire : en fait, on peut montrer que la "bonne" mesure de probabilité $\mathbb{P}_2$ est donnée par

$$\mathbb{P}_2(\{i,j\}) = \begin{cases} 1/18 & \text{si } i \neq j, \\ 1/36 & \text{si } i = j, \end{cases} \quad (i,j) \in \Omega_2. \tag{1.3}$$

De même, si l'on décrit cette expérience à l'aide de l'univers Ω_3, ce qui revient à enregistrer le plus petit résultat et la différence entre le plus grand et le plus petit résultat, la "bonne" mesure de probabilité $\mathbb{P}_3$ est donnée par

$$\mathbb{P}_3(\{i,j\}) = \begin{cases} 1/18 & \text{si } j \neq 0, \\ 1/36 & \text{si } j = 0, \end{cases} \quad (i,j) \in \Omega_3. \tag{1.4}$$

Ainsi, les trois espaces de probabilité $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ pour $i = 1, 2, 3$ sont différents mais décrivent la même expérience aléatoire.

Proposition 1.1.1. *Soit $(A_n, n \in \mathbb{N})$ une suite d'évènements, i.e., $A_n \in \mathcal{F}$ pour $n \in \mathbb{N}$. Toute mesure de probabilité $\mathbb{P}$ sur $(\Omega, \mathcal{F})$ satisfait les propriétés suivantes :*

- $\mathbb{P}(A_1^c) = 1 - \mathbb{P}(A_1)$;
- $\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2)$;
- *si la suite $(A_n, n \in \mathbb{N})$ est croissante, i.e., $A_n \subset A_{n+1}$, alors*

$$\mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n);$$

- *si la suite $(A_n, n \in \mathbb{N})$ est décroissante, i.e., $A_{n+1} \subset A_n$, alors*

$$\mathbb{P}\left(\bigcap_{n\in\mathbb{N}} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

1.2 Variables aléatoires discrètes

1.2.1 Définition et loi d'une variable aléatoire discrète

Lorsque Ω est dénombrable, comme c'est le cas dans tout ce chapitre, la définition d'une **variable aléatoire** est particulièrement simple. Dans le cas où Ω n'est pas dénombrable il faudra prendre quelques précautions, cf. Section 2.5.

Définition 1.2.1. Soit Ω' un ensemble quelconque (en particulier, dénombrable ou non) et $\mathcal{F}' = \mathcal{P}(\Omega')$ l'ensemble de ses parties. Une application $X : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$ est appelée **variable aléatoire discrète**.
On dit que X est **à valeurs dans** $\mathbb{R}$ (ou $\mathbb{N}$, $\mathbb{R}^n$, etc) si $X(\Omega) \subset \mathbb{R}$ (ou $\mathbb{N}$, $\mathbb{R}^n$, etc). Lorsque X est à valeurs dans $\mathbb{R}^n$ avec $n \geq 2$, on dit que X est un **vecteur aléatoire discret**.

Dans le reste de ce chapitre, en plus de considérer Ω est dénombrable et que $\mathcal{F} = \mathcal{P}(\Omega)$ on utilisera les notations suivantes :

- Ω', Ω'', Ω_1, Ω_2, etc, sont des ensembles arbitraires ;
- $\mathcal{F}' = \mathcal{P}(\Omega')$, $\mathcal{F}'' = \mathcal{P}(\Omega'')$, $\mathcal{F}_1 = \mathcal{P}(\Omega_1)$, $\mathcal{F}_2 = \mathcal{P}(\Omega_2)$, etc ;
- X, X', X'', Y, X_1, X_2, etc, sont des variables aléatoires définies sur $(\Omega, \mathcal{F})$ et à valeurs dans des ensembles arbitraires que l'on spécifiera le cas échéant.

A ce stade il peut paraître mystérieux de faire apparaître les ensembles $\mathcal{F}$ et $\mathcal{F}'$ dans la définition d'une variable aléatoire puisqu'ils ne jouent apparemment aucun rôle, mais cela prendra tout son sens lorsque l'on parlera de variables aléatoires absolument continues au Chapitre 2. A cause de la définition très générale d'une variable aléatoire discrète, le fait d'être une variable aléatoire discrète est une propriété extrêmement stable. On notera par exemple les propriétés suivantes.

Proposition 1.2.1. *Pour toute application $\varphi : (\Omega', \mathcal{F}') \to (\Omega'', \mathcal{F}'')$, l'application $\varphi \circ X : (\Omega, \mathcal{F}) \to (\Omega'', \mathcal{F}'')$ est une variable aléatoire discrète.*
Ainsi, si X_1 et X_2 sont deux variables aléatoires à valeurs dans $\Omega' = \mathbb{R}$, alors toute combinaison linéaire de X_1 et X_2 est une variable aléatoire, ainsi que X_1^2, $\min(X_1, X_2)$, etc.

Les variables aléatoires sont des objets très importants puisqu'ils capturent ce qui, dans l'expérience aléatoire, nous intéresse vraiment. Ainsi, pour revenir à l'exemple 1.1.2 du lancer de deux dés indiscernables, ce qui nous intéresse parfois n'est pas tant le résultat de chaque dé mais plutôt leur somme : dans ce cas, on considèrera la variable aléatoire X_2 ou X_3 définie par

$$X_2 : (i,j) \in \Omega = \Omega_2 \mapsto i+j \in \Omega' = \{2,\dots,12\} \tag{1.5}$$

ou bien par

$$X_3 : (i,j) \in \Omega = \Omega_3 \mapsto 2i+j \in \Omega' = \{2,\dots,12\}, \tag{1.6}$$

suivant la modélisation de l'expérience adoptée. Par la suite, la question naturelle est de connaître la probabilité d'obtenir un certain résultat, par exemple, quelle est la probabilité d'obtenir un résultat ≥ 5 ? Formellement et plus généralement, il s'agit de **calculer la loi de** X. Pour définir cette notion, on rappelle la définition ensembliste de l'**image réciproque** d'une fonction : si $f : \Omega \to \Omega'$ est une fonction quelconque, son image réciproque $f^{-1} : \mathcal{P}(\Omega') \to \mathcal{P}(\Omega)$ est la fonction définie par

$$f^{-1}(A) = \{\omega \in \Omega : f(\omega) \in A\}\,,\ A \subset \Omega'.$$

Lorsque f est bijective, l'image réciproque de tout évènement élémentaire est un singleton, disons $f^{-1}(\{\omega'\}) = \{\omega\}$, et par abus de notation on identifiera alors f^{-1} avec l'inverse de f, définie dans l'exemple précédent par la relation $f^{-1}(\omega') = \omega$.

Définition 1.2.2. La **loi de la variable aléatoire** X est l'application $\mathbb{P}\circ X^{-1} : \mathcal{F}' \to [0,1]$.

De manière plus légère, la loi de X sera parfois notée $\mathbb{P}_X = \mathbb{P}\circ X^{-1}$. Il s'agit donc de la fonction qui, à un évènement $A \in \mathcal{F}'$, associe le nombre

$$\mathbb{P}_X(A) = \mathbb{P}\circ X^{-1}(A) = \mathbb{P}\left(X^{-1}(A)\right) = \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \in A\}\right) \in [0,1].$$

Remarque 1.2.1. La loi de X dépend de la mesure de probabilité considérée (ici, $\mathbb{P}$). Si l'on est amenés à considérer plusieurs mesures de probabilités (ce qui sera le cas lorsque l'on fera de la statistique), on pourra alors être amené à parler de la **loi de** X **sous** $\mathbb{P}$.

Théorème 1.2.2. *Si $X : (\Omega,\mathcal{F}) \to (\Omega',\mathcal{F}')$, alors la loi de X est une mesure de probabilité sur $(\Omega',\mathcal{F}')$.*

En d'autres termes,

Toute variable aléatoire X induit
une mesure de probabilité sur son ensemble image, appelée loi de X.

Démonstration du théorème 1.2.2. Il s'agit de vérifier les deux axiomes d'une mesure de probabilité. Tout d'abord, on calcule bien $\mathbb{P}_X(\Omega') = 1$ puisque $\mathbb{P}_X(\Omega') = \mathbb{P}(X^{-1}(\Omega'))$ par définition de $\mathbb{P}_X$, que $X^{-1}(\Omega') = \Omega$ par définition de X^{-1} et finalement que $\mathbb{P}(\Omega) = 1$ par hypothèse. La propriété (1.1) suit du fait que l'image réciproque commute avec l'union, i.e.,

$$X^{-1}\left(\bigcup_{i\in I} A_i\right) = \bigcup_{i\in I} X^{-1}\left(A_i\right),$$

qui implique que $\mathbb{P}_X(\cup_{i\in I}A_i) = \mathbb{P}(\cup_{i\in I}X^{-1}(A_i))$, et du fait que si (A_i) est une famille d'ensembles deux à deux disjoints, alors la famille $(X^{-1}(A_i))$ l'est aussi, ce qui implique que $\mathbb{P}(\cup_{i\in I}X^{-1}(A_i)) = \sum_{i\in I}\mathbb{P}(X^{-1}(A_i)) = \sum_{i\in I}\mathbb{P}_X(A_i)$. ■

Deux variables aléatoires différentes peuvent avoir la même loi. Par exemple, les variables aléatoires X_2 et X_3 de (1.5) et (1.6) décrivent le même résultat d'une même expérience aléatoire, à savoir la somme de deux dés. Il est donc évident que les lois de X_2 et X_3 doivent coïncider : la probabilité d'obtenir 2 ne doit pas dépendre du formalisme adopté ! Ainsi, si les mesures de probabilité $\mathbb{P}_2$ et $\mathbb{P}_3$ de (1.3) et (1.4) ont été bien choisis, on doit avoir

$$\mathbb{P}_2 \circ X_2^{-1} = \mathbb{P}_3 \circ X_3^{-1}$$

ce que l'on peut effectivement vérifier analytiquement.

1.2.2 Fonctions indicatrices

Une famille de variables aléatoires discrètes joue un rôle très important en théorie des probabilités : il s'agit des fonctions indicatrices.

Définition 1.2.3. Soit $A \in \mathcal{F}$. La **fonction indicatrice de l'évènement** A, notée ξ_A, est la variable aléatoires à valeurs dans $\{0, 1\}$ définie de la manière suivante : $\xi_A = 1$ si l'évènement A a lieu et 0 sinon, i.e.,

$$\xi_A(\omega) = \begin{cases} 1 & \text{si } \omega \in A, \\ 0 & \text{sinon.} \end{cases}$$

On notera $\mathbb{1}\{X \in A\}$ la fonction indicatrice de l'évènement $X^{-1}(A)$, i.e.,

$$\mathbb{1}\{X \in A\}(\omega) = \mathbb{1}\{X(\omega) \in A\} = \begin{cases} 1 & \text{si } X(\omega) \in A, \\ 0 & \text{sinon.} \end{cases}$$

1.2.3 La disparition, ou bien le formalisme probabiliste

Le fait qu'une même expérience aléatoire puisse être décrite par plusieurs espaces de probabilité montre que le choix de l'espace de probabilité n'est pas très important. En pratique, on est souvent beaucoup plus intéressés par les variables aléatoires et leur loi, les lois étant, comme on l'a vu, indépendantes de l'espace de probabilité choisi. Ainsi, le formalisme probabiliste vise à faire disparaître l'espace de probabilité sous-jacent pour mettre l'accent sur les variables aléatoires et leur loi et au lieu de noter

$$\mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\}),$$

ce qui serait parfaitement rigoureux, on notera de manière plus légère

$$\mathbb{P}(X \in A)$$

qui s'interprète tout simplement par la probabilité que X appartienne à A. Il s'agit du premier exemple où Ω, bien qu'indispensable pour la construction rigoureuse de la théorie des probabilités, va peu à peu s'effacer au profit des lois induites par les variables aléatoires considérées. De la même manière, au lieu de noter $\varphi \circ X$ on notera tout simplement $\varphi(X)$, etc.

Remarque 1.2.2. Ce formalisme peut parfois prêter à confusion. Considérons par exemple le raisonnement suivant : soit X une variable aléatoire discrète à valeurs réelles et $F_X : x \in \mathbb{R} \to \mathbb{P}(X \leq x)[0, 1]$ (comme on le verra plus loin, F_X est la fonction de répartition de X). Une interprétation incorrecte du formalisme probabiliste pourrait alors mener à la conclusion que $F_X(X) = 1$: en effet, puisque $F_X(x) = \mathbb{P}(X \leq x)$ on a $F_X(X) = \mathbb{P}(X \leq X) = 1$. Ce raisonnement est bien entendu erroné, mais saurez-vous trouver le problème ?

1.2.4 Quelques lois discrètes classiques

Une liste non-exhaustive de lois absolument continues inclut :

Loi uniforme sur un ensemble fini : Soit A un ensemble fini : la loi uniforme sur A, mentionnée précédemment, est la mesure de probabilité $\mathbb{P}(\{a\}) = \frac{1}{|A|}$ pour $a \in A$. Elle représente l'expérience où l'on tire un élément uniformément au hasard d'un ensemble fini ;

Loi de Bernoulli : La loi de Bernoulli est la mesure de probabilité la plus simple possible : c'est une mesure de probabilité sur $\{0, 1\}$, est elle donc entièrement caractérisée par la probabilité de 1 : $\mathbb{P}(\{1\}) = p \in [0, 1]$. La loi de Bernoulli est typiquement la loi du pile ou face ;

Loi binomiale : La loi binomiale est une loi sur $\mathbb{N}$ avec deux paramètres : $p \in]0, 1[$ et $n \in \mathbb{N}^*$. La loi binomiale est la loi de la variable aléatoire qui compte le nombre de succès dans une série de n lancers indépendants lorsque chaque succès a lieu avec probabilité p.

$$\mathbb{P}(\{k\}) = \binom{n}{k} p^k (1-p)^{n-k}, \ k = 0, \dots, n;$$

Loi géométrique : La loi géométrique est la loi du nombre de lancers jusqu'au premier succès, lorsque les lancers sont indépendants et ont succès avec probabilité $p \in]0, 1[$: ainsi, c'est une mesure de probabilité sur $\mathbb{N}^*$ paramétrée par un nombre réel $p \in]0, 1[$ et donnée par

$$\mathbb{P}(\{k\}) = (1-p)^{k-1} p, \ k \in \mathbb{N}^*;$$

Loi de Poisson : La loi de Poisson, caractérisée par un paramètre $\lambda \in > 0$, est la mesure de probabilité sur $\mathbb{N}$ définie par

$$\mathbb{P}(\{k\}) = \frac{\lambda^k}{k!} e^{-\lambda}, \ k \in \mathbb{N}.$$

Il s'agit d'une loi essentielle pour la modélisation des évènements rares qui émerge comme limite de la loi binomiale, cf. Proposition 3.4.5.

1.3 Variables aléatoires discrètes à valeurs réelles : fonction de répartition, espérance et variance

Dans toute cette partie (Section 1.3), **on suppose que X est à valeurs réelles** et l'on introduit plusieurs notions importantes. Ces notions seront généralisées dans la Section 1.4 au cas de vecteurs aléatoires discrets (toujours à valeurs réelles).

1.3.1 Fonction de répartition

Définition 1.3.1. La **fonction de répartition** de la variable aléatoire X à valeurs réelles est la fonction $F_X : \mathbb{R} \to [0, 1]$ définie par $F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}_X((-\infty, x])$ pour $x \in \mathbb{R}$.

Les propriétés suivantes découlent directement de cette définition et des propriétés d'une loi de probabilité établie dans la Proposition 1.1.1.

Proposition 1.3.1. *La fonction F_X est croissante, continue à droite et satisfait*

$$\lim_{x\to-\infty} F_X(x) = 0, \quad \lim_{x\to+\infty} F_X(x) = 1 \quad \textit{et} \quad F_X(x) - F_X(x-) = \mathbb{P}(X = x) = \mathbb{P}_X(\{x\}), \ x \in \mathbb{R}.$$

Corollaire 1.3.2. *La fonction de répartition caractérise la loi d'une variable aléatoire : si $F_X = F_Y$ alors $\mathbb{P}_X = \mathbb{P}_Y$.*

Démonstration. Cela vient directement du fait que $F_X(x) - F_X(x-) = \mathbb{P}_X(\{x\})$ et du fait qu'une mesure de probabilité discrète est caractérisée par les probabilités des évènements élémentaires. ■

1.3.2 Espérance

Pour $x \in \mathbb{R}$ on rappelle les notations

$$x^+ = \max(x, 0) = \begin{cases} x & \text{si } x \geq 0 \\ 0 & \text{sinon} \end{cases} \quad \text{et} \quad x^- = -\min(x, 0) = \begin{cases} -x & \text{si } x \leq 0 \\ 0 & \text{sinon} \end{cases}.$$

Définition 1.3.2. Si X est à valeurs dans $\mathbb{R}_+ \cup \{\infty\}$, alors son **espérance** $\mathbb{E}(X) \in \mathbb{R}_+ \cup \{\infty\}$ est définie de la manière suivante :

$$\mathbb{E}(X) = \sum_{x\in X(\Omega)} x\mathbb{P}_X(\{x\}) = \sum_{x\in X(\Omega)} x\mathbb{P}(X = x) \tag{1.7}$$

avec la convention $\mathbb{E}(X) = \infty$ si $\mathbb{P}(X = \infty) > 0$.

Si X est à valeurs dans $\mathbb{R}$, X est dite **intégrable** si $\mathbb{E}(X^+), \mathbb{E}(X^-) < \infty$. Dans ce cas, son **espérance** $\mathbb{E}(X) \in \mathbb{R}$ est donnée par $\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-)$.

Si X est à valeurs dans $\mathbb{R}$, X est dite de **carré intégrable** si X^2 est intégrable, i.e., $\mathbb{E}(X^2) < \infty$.

Remarque 1.3.1. Formellement, $\mathbb{E}$ est un opérateur qui agit sur les variables aléatoires et qui dépend de la mesure de probabilité considérée, ici, $\mathbb{P}$. Ainsi, chaque nouvelle mesure de probabilité induit un nouvel opérateur d'espérance associé. Dans le cadre de ce cours il y aura notamment deux cas particuliers importants :

- comme on le verra dans la Section 1.6, chaque évènement $A \in \mathcal{F}$ avec $\mathbb{P}(A) > 0$ induit une nouvelle mesure de probabilité $\mathbb{P}(\cdot \mid A)$, appelée mesure de probabilité conditionnelle sachant A, et l'opérateur associé est $\mathbb{E}(\cdot \mid A)$;
- en statistique, on considérera par exemple une famille $\{\mathbb{P}_\theta, \theta \in \Theta\}$ de mesures de probabilités sur un même espace, et $\mathbb{E}_\theta$ dénotera alors l'espérance associée à la mesure de probabilité $\mathbb{P}_\theta$.

Remarque 1.3.2. L'espérance de X ne dépend que de sa loi : deux variables aléatoires avec la même loi ont donc la même espérance.

Nous montrons maintenant quelques propriétés essentielles de l'espérance :

- la Proposition 1.3.5 donne une formule explicite pour $\mathbb{E}(|X|)$ de laquelle découle que la formule (1.7) reste valable dans le cas où X est intégrable ;
- le Théorème 1.3.6 qui établit une formule fondamentale pour calculer l'espérance de $\varphi(X)$;

- la Proposition 1.3.7 qui montre que l'espérance est un opérateur linéaire ;
- le Théorème 1.3.13 qui montre l'écriture probabiliste de l'inégalité de Cauchy–Schwarz.

Les preuves de ce résultat sont élémentaires, et reposent essentiellement sur le théorème de Fubini et les deux résultats suivants. Le premier résultat montre un lien profond entre probabilité et espérance : $\mathbb{P}(A)$ n'est rien d'autre que l'espérance de la variable aléatoire ξ_A.

Théorème 1.3.3. *Pour tout $A \in \mathcal{F}$ on a $\mathbb{E}(\xi_A) = \mathbb{P}(A)$.*
En particulier, pour tout $A' \in \mathcal{F}'$ on a $\mathbb{E}\left(\mathbb{1}\left\{X \in A'\right\}\right) = \mathbb{P}(X \in A')$.

Démonstration. Il suffit juste de montrer la première propriété puisque la deuxième propriété en découle en considérant $A = X^{-1}(A')$. Puisque $\xi_A \subset \{0,1\}$, on a par définition de l'espérance

$$\mathbb{E}(\xi_A) = \sum_{y \in \{0,1\}} y\mathbb{P}(\xi_A = y) = \mathbb{P}(\xi_A = 1).$$

Par définition de ξ_A, $\mathbb{P}(\xi_A = 1) = \mathbb{P}(A)$ ce qui prouve le résultat voulu. ■

Proposition 1.3.4. *Si X est à valeurs dans $\mathbb{R}_+$,*

$$\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X \geq x)\mathrm{d}x.$$

Démonstration. Pour $x \in \mathbb{R}_+$, on a $x = \int \mathbb{1}\left\{0 \leq u \leq x\right\} \mathrm{d}u$: ainsi, puisque $X(\Omega) \subset \mathbb{R}_+$ par hypothèse, cela nous donne

$$\mathbb{E}(X) = \sum_{x \in X(\Omega)} x\mathbb{P}_X(\{x\}) = \sum_{x \in X(\Omega)} \int \mathbb{1}\left\{0 \leq u \leq x\right\} \mathbb{P}_X(\{x\})\mathrm{d}u.$$

Puisque tous les termes sont positifs, le théorème de Fubini nous permet d'intervertir somme et intégrale ce qui nous donne, en intégrant d'abord sur $u \geq 0$,

$$\mathbb{E}(X) = \int \mathrm{d}u \mathbb{1}\left\{u \geq 0\right\} \left(\sum_{x \in X(\Omega)} \mathbb{1}\left\{x \geq u\right\} \mathbb{P}_X(\{x\}) \right).$$

Par définition, $\sum_{x \in X(\Omega)} \mathbb{1}\left\{x \geq u\right\} \mathbb{P}_X(\{x\}) = \mathbb{P}(X \geq u)$ et l'on obtient donc le résultat. ■

Proposition 1.3.5. *Soit X à valeurs dans $\mathbb{R}$: alors*

$$\mathbb{E}\left(|X|\right) = \mathbb{E}\left(X^+\right) + \mathbb{E}\left(X^-\right) = \sum_{x \in X(\Omega)} |x|\mathbb{P}_X(\{x\}).$$

En particulier, on a les équivalences suivantes :

$$X \text{ est intégrable } \iff \mathbb{E}(|X|) < \infty \iff \sum_{x \in X(\Omega)} |x|\mathbb{P}_X(\{x\}) < \infty.$$

Si l'une de ces trois conditions est satisfaite, alors la série $(x\mathbb{P}_X(\{x\}), x \in X(\Omega))$ est sommable et la formule (1.7) reste valable, i.e.,

$$\mathbb{E}(X) = \sum_{x \in X(\Omega)} x\mathbb{P}_X(\{x\}) = \sum_{x \in X(\Omega)} x\mathbb{P}(X = x).$$

Enfin, si X est intégrable alors $|\mathbb{E}(X)| \leq \mathbb{E}(|X|)$.

Démonstration. On vérifie que pour $x > 0$ les évènements $\{X^+ \geq x\}$ et $\{X^- \geq x\}$ sont disjoints et que leur union est égale à $\{|X| \geq x\}$: ainsi, la Proposition 1.1.1 implique que $\mathbb{P}(|X| \geq x) = \mathbb{P}(X^+ \geq x) + \mathbb{P}(X^- \geq x)$ et il s'ensuit grâce à la Proposition 1.3.4 que

$$\mathbb{E}\left(|X|\right) = \int_0^\infty \mathbb{P}\left(X^+ \geq x\right) \mathrm{d}x + \int_0^\infty \mathbb{P}\left(X^- \geq x\right) \mathrm{d}x = \mathbb{E}\left(X^+\right) + \mathbb{E}\left(X^-\right).$$

Par ailleurs, on vérifie que

$$\mathbb{E}\left(X^\pm\right) = \sum_{x \in X^\pm(\Omega)} x \mathbb{P}_{X^\pm}(\{x\}) = \sum_{x \in X(\Omega)} x^\pm \mathbb{P}_X(\{x\})$$

si bien que

$$\mathbb{E}(X^+) + \mathbb{E}(X^-) = \sum_{x \in X(\Omega)} (x^+ + x^-) \mathbb{P}_X(\{x\}) = \sum_{x \in X(\Omega)} |x| \mathbb{P}_X(\{x\}).$$

On a donc montré que $\mathbb{E}(|X|) = \sum_{x \in X(\Omega)} |x| \mathbb{P}_X(\{x\})$, d'où le reste de la proposition découle facilement. ■

Théorème 1.3.6. *Soit $\varphi : \Omega' \to \mathbb{R}$: si $\varphi(X) \geq 0$ ou $\mathbb{E}(|\varphi(X)|) < \infty$, alors*

$$\mathbb{E}\left(\varphi(X)\right) = \sum_{x \in X(\Omega)} \varphi(x) \mathbb{P}_X(\{x\}) = \sum_{x \in X(\Omega)} \varphi(x) \mathbb{P}(X = x). \tag{1.8}$$

Démonstration. On traite le cas $\varphi(X) \geq 0$, duquel le cas intégrable suit en revenant à la définition de $\mathbb{E}(X)$ pour X intégrable. Pour $\varphi(X) \geq 0$, on peut utiliser la Proposition 1.3.4 qui donne

$$\mathbb{E}(\varphi(X)) = \int \mathbb{P}(\varphi(X) \geq x) \mathbb{1}\left\{x \geq 0\right\} \mathrm{d}x = \int \mathbb{P}\left(X \in \varphi^{-1}([x, \infty))\right) \mathbb{1}\left\{x \geq 0\right\} \mathrm{d}x.$$

Pour n'importe quel $A \in \mathcal{F}$, on a $\mathbb{P}(X \in A) = \sum_{x \in X(\Omega) \cap A} \mathbb{P}_X(\{x\})$ du fait que $\{X \in A\} = \cup_{x \in X(\Omega) \cap A} \{X = x\}$, d'où il vient que

$$\mathbb{P}\left(X \in \varphi^{-1}([x, \infty))\right) = \sum_{u \in X(\Omega)} \mathbb{1}\left\{u \in \varphi^{-1}([x, \infty))\right\} \mathbb{P}_X(\{u\}) = \sum_{u \in X(\Omega)} \mathbb{1}\left\{\varphi(u) \geq x\right\} \mathbb{P}_X(\{u\}).$$

Ainsi,

$$\mathbb{E}\left(\varphi(X)\right) = \int \sum_{u \in X(\Omega)} \mathbb{1}\left\{\varphi(u) \geq x \geq 0\right\} \mathbb{P}_X(\{u\}) \mathrm{d}x.$$

Puisque tous les termes sont positifs, le théorème de Fubini nous garantit qu'on peut intervertir somme et intégrale : intégrant d'abord (sur x), on obtient alors le résultat puisque $\varphi(u) = \int \mathbb{1}\left\{0 \leq x \leq \varphi(u)\right\} \mathrm{d}x$. ■

Proposition 1.3.7. *Soit Y une variable aléatoire à valeurs réelles. Si $X, Y \in \mathbb{R}_+$, alors pour tout $a, b \in \mathbb{R}_+$ on a $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$. Cette relation reste valable pour tout $a, b \in \mathbb{R}$ dès lors que $X, Y \in \mathbb{R}$ sont intégrables. En particulier, si X et Y sont intégrables et que $X \geq Y$, alors $\mathbb{E}(X) \geq \mathbb{E}(Y)$.*

Démonstration. Soit $a, b \geq 0$ et X, Y à valeurs dans $\mathbb{R}_+$: alors l'égalité $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$ découle du Théorème 1.3.6 avec $X = (X, Y)$ et $\varphi(x, y) = ax + by$. En outre, si $Y \geq X$ alors $\mathbb{P}(X \geq x) = \mathbb{P}(Y \geq X \geq x) \leq \mathbb{P}(Y \geq x)$ et donc (en utilisant la Proposition 1.3.4 pour calculer les espérances)

$$\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X \geq x)\mathrm{d}x \leq \int_0^\infty \mathbb{P}(Y \geq x)\mathrm{d}x = \mathbb{E}(Y)$$

ce qui prouve le résultat dans ce cas. Considérons maintenant le cas général. Alors l'inégalité triangulaire donne $|aX + bY| \leq |a||X| + |b||Y|$ et donc, d'après la première étape,

$$\mathbb{E}\left(|aX + bY|\right) \leq |a|\mathbb{E}\left(|X|\right) + |b|\mathbb{E}\left(|Y|\right).$$

Ainsi, si X et Y sont intégrables $aX + bY$ l'est aussi et l'on retrouve alors la formule $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$ comme dans le cas positif en invoquant le Théorème 1.3.6. Enfin, si $Y \geq X$ alors $Y - X \geq 0$ et donc $\mathbb{E}(Y - X) \geq 0$ ce qui donne le résultat puisque $\mathbb{E}(Y - X) = \mathbb{E}(Y) - \mathbb{E}(X)$ lorsque X et Y sont intégrables, comme l'on vient de le voir. ■

Le résultat précédent se généralise de la manière suivante.

Proposition 1.3.8. *Si $X_i \geq 0$, alors*

$$\mathbb{E}\left(\sum_{i\geq 1} X_i\right) = \sum_{i\geq 1} \mathbb{E}(X_i).$$

1.3.3 Variance et écart-type

Définition 1.3.3. Si X à valeurs réelles est intégrable, alors sa **variance** $\mathbb{V}\mathrm{ar}(X) \in \mathbb{R}_+ \cup \{\infty\}$ et son **écart-type** $\sigma_X \in \mathbb{R}_+ \cup \{\infty\}$ sont définis de la manière suivante :

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] \quad \text{et} \quad \sigma_X = \sqrt{\mathbb{V}\mathrm{ar}(X)}.$$

La variance mesure la dispersion d'une variable aléatoire autour de sa moyenne, ce qui est par exemple reflété par l'inégalité de Bienaymé–Tchebychev 1.3.12 ci-dessous.

Proposition 1.3.9. *Si X est intégrable, alors pour tout $a, b \in \mathbb{R}$ on a*

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}(X^2) - \big(\mathbb{E}(X)\big)^2 \quad \textit{et} \quad \mathbb{V}\mathrm{ar}(aX + b) = a^2\mathbb{V}\mathrm{ar}(X).$$

Ainsi, $\mathbb{V}\mathrm{ar}(X) < \infty$ si et seulement si X est de carré intégrable.

Le résultat suivant, qui permet de caractériser le fait qu'une variable aléatoire est constante, est une conséquence directe de l'inégalité de Bienaymé–Tchebychev ci-dessous (Théorème 1.3.12).

Proposition 1.3.10. *Si X est intégrable et que $\mathbb{V}\mathrm{ar}(X) = 0$, alors X est (presque sûrement) constante, i.e., $\mathbb{P}(X = \mathbb{E}(X)) = 1$.*

Démonstration. On a $\mathbb{P}(X \neq \mathbb{E}(X)) = \lim_{\varepsilon \downarrow 0} \mathbb{P}(|X - \mathbb{E}(X)| \geq \varepsilon)$ et pour tout $\varepsilon > 0$, l'inégalité de Bienaymé–Tchebychev donne $\mathbb{P}(|X - \mathbb{E}(X)| \geq \varepsilon) = 0$. ■

1.3.4 Inégalités de Markov, de Cauchy–Schwarz et de Bienaymé–Tchebychev

On formule maintenant trois inégalités probabilistes classiques. Bien qu'apparemment élémentaire, les conséquences de l'inégalité de Markov sont très riches.

Théorème 1.3.11 (Inégalité de Markov). *Soit X à valeurs dans $\mathbb{R}_+$: alors pour tout $x > 0$,*

$$\mathbb{P}(X \geq x) \leq \frac{1}{x}\mathbb{E}(X).$$

Démonstration. Il suffit de remarquer que

$$\mathbb{1}\{X \geq x\} \leq \frac{X}{x}$$

puis de prendre l'espérance. ■

Théorème 1.3.12 (Inégalité de Bienaymé–Tchebychev). *Si X à valeurs réelles est intégrable, alors pour tout $\varepsilon > 0$ on a*

$$\mathbb{P}\left(|X - \mathbb{E}(X)| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2}\mathbb{V}\mathrm{ar}(X).$$

Démonstration. On écrit

$$\mathbb{P}\left(|X - \mathbb{E}(X)| \geq \varepsilon\right) = \mathbb{P}\left((X - \mathbb{E}(X))^2 \geq \varepsilon^2\right)$$

puis on utilise l'inégalité de Markov. ■

Théorème 1.3.13 (Inégalité de Cauchy–Schwarz). *On a $\mathbb{E}(|XY|) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}$. En particulier, si X est de carré intégrable alors X est intégrable.*

Démonstration. La Proposition 1.3.5 donne $\mathbb{E}(|XY|) = \sum_{x\in X(\Omega), y\in Y(\Omega)} |xy| \mathbb{P}_{X,Y}(\{x,y\})$ et donc l'inégalité de Cauchy–Schwarz donne

$$\mathbb{E}(|XY|) \leq \left(\sum_{\substack{x\in X(\Omega)\\ y\in Y(\Omega)}} \left(|x|\mathbb{P}_{X,Y}(\{x,y\})^{1/2}\right)^2\right)^{1/2} \left(\sum_{\substack{x\in X(\Omega)\\ y\in Y(\Omega)}} \left(y\mathbb{P}_{X,Y}(\{x,y\})^{1/2}\right)^2\right)^{1/2}.$$

Par ailleurs,

$$\begin{aligned}\sum_{\substack{x\in X(\Omega)\\ y\in Y(\Omega)}} \left(|x|\mathbb{P}_{X,Y}(\{x,y\})^{1/2}\right)^2 &= \sum_{\substack{x\in X(\Omega)\\ y\in Y(\Omega)}} x^2 \mathbb{P}_{X,Y}(\{x,y\}) \\ &= \sum_{x\in X(\Omega)} x^2 \sum_{y\in Y(\Omega)} \mathbb{P}_{X,Y}(\{x,y\}) \\ &= \sum_{x\in X(\Omega)} x^2 \mathbb{P}_X(\{x\})\end{aligned}$$

qui vaut $\mathbb{E}(X^2)$ par le Théorème 1.3.6 pour $\varphi(x) = x^2$, et où on a utilisé le Théorème 1.4.1 pour la troisième égalité ci-dessus. ■

La réciproque de la propriété précédente n'est pas vraie, i.e., on peut avoir une variable aléatoire intégrable mais pas de carré intégrable : il suffit par exemple de considérer X avec

$$\mathbb{P}(X = x) = \frac{1}{Zx^{5/2}},\ x \in \mathbb{N}^*, \quad \text{où} \quad Z = \sum_{x\in\mathbb{N}^*} \frac{1}{x^{5/2}}.$$

1.4 Famille de variables aléatoires discrètes, indépendance

1.4.1 Famille de variables aléatoires discrètes : définition et lois marginales

En pratique, il est très utile de considérer des familles de variables aléatoires discrètes. Reprenons l'exemple 1.1.6 du lancer de deux dés discernables décrit par l'espace de probabilité $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ et définissons $X_1 : (i,j) \in \Omega_1 \mapsto i \in \{1,\dots,6\}$ et $X_2 : (i,j) \in \Omega_1 \mapsto j \in \{1,\dots,6\}$. Ainsi, X_i est la variable aléatoire qui enregistre le résultat du i-ème dé, et on dit que $X = (X_1, X_2)$ est un **couple de variables aléatoires discrètes**, ou tout simplement **couple aléatoire discret**. De manière plus générale, si chaque X_k pour $k \in \mathbb{N}^*$ est une variable aléatoire discrète alors $(X_1,\dots,X_n)$ est appelé **vecteur aléatoire discret** et $(X_n, n \in \mathbb{N}^*)$ est appelé **famille aléatoire discrète**. On rappelle que par définition, $x \le y$ pour $x, y \in \mathbb{R}^n$ si et seulement si $x_k \le y_k$ pour $k = 1,\dots,n$, ce qui permet d'étendre la Définition 1.3.1 de la fonction de répartition.

Définition 1.4.1. La **fonction de répartition** de la variable aléatoire X à valeurs dans $\mathbb{R}^n$ est la fonction $F_X : \mathbb{R}^n \to [0,1]$ définie par $F_X(x) = \mathbb{P}(X \le x)$ pour $x \in \mathbb{R}^n$.

Le Corollaire 1.3.2 continue à être vrai dans le cadre vectoriel, i.e., la fonction de répartition caractérise la loi d'un vecteur aléatoire discret à valeurs réelles. Dans l'exemple ci-dessus, l'intérêt de considérer la variable aléatoire $X = (X_1, X_2)$ est que l'on peut par exemple utiliser le Théorème 1.3.6 pour calculer l'espérance d'autres variables aléatoires d'intérêt. Si l'on s'intéresse par exemple au maximum des deux dés, il suffit de prendre $\varphi(x,y) = \max(x,y)$ puis de calculer

$$\mathbb{E}(\varphi(X)) = \mathbb{E}(\max(X_1,X_2)) = \sum_{x_1,x_2=1,\dots,6} \max(x_1,x_2)\mathbb{P}_1(X_1 = x_1, X_2 = x_2).$$

Puisque tous les résultats sont équiprobables, $\mathbb{P}_1$ est la mesure uniforme sur Ω_1 et donc

$$\mathbb{E}(\varphi(X)) = \frac{1}{36} \sum_{x_1,x_2=1,\dots,6} \max(x_1,x_2) = \frac{161}{36}.$$

De manière plus générale, si $(X_n, n \in \mathbb{N})$ est une famille aléatoire discrète, il y a deux manières de considérer la variable aléatoire discrète X_n et sa loi :

1. soit on considère X_n en isolation, ce qui correspond à ce qu'on a fait jusqu'à présent ;
2. soit on considère X_n comme la projection sur la coordonnée n de la famille $(X_n, n \in \mathbb{N})$: dans ce cas et pour mettre l'accent sur le fait que X_n fait partie d'une famille plus grande, la loi de X_n sera appelée **loi marginale de** X_n.

Le Théorème 1.3.6 utilisé avec la fonction $f((x_n, n \in \mathbb{N})) = x_n$, i.e., la projection sur la coordonnée n, permet de calculer la loi marginale de X_n.

Théorème 1.4.1. *Soit $X = (X_1, \dots, X_n)$ un vecteur aléatoire discret avec $X_k : (\Omega, \mathcal{F}) \to (\Omega_k, \mathcal{F}_k)$ pour $k = 1, \dots, n$. Alors pour tout $x_1 \in X_1(\Omega)$, la loi marginale de X_1 est donnée par*

$$\mathbb{P}_{X_1}(\{x_1\}) = \sum_{x_2 \in X_2(\Omega), \dots, x_n \in X_n(\Omega)} \mathbb{P}_X\left(\{(x_1, x_2, \dots, x_n)\}\right).$$

Démonstration. Il suffit, pour $x_1 \in X_1(\Omega)$ fixé, d'utiliser le Théorème 1.3.6 avec la fonction $\varphi(y_1, \dots, y_n) = \mathbb{1}\{y_1 = x_1\}$. ■

Dans le cas d'un couple aléatoire discret (X_1, X_2) à valeurs réelles, on a par exemple

$$\mathbb{P}_{X_1}(A) = \sum_{\substack{x_1 \in A \\ x_2 \in X_2(\Omega)}} \mathbb{P}\left(X_1 = x_1, X_2 = x_2\right), \ A \subset \mathbb{R}.$$

Cet exemple est en fait le plus général possible, puisqu'on peut s'y ramener en définissant $X_2 = (X_2, \dots, X_n)$ dans le théorème précédent. On retiendra donc que

La loi marginale de X_1 est obtenue en sommant sur les valeurs possibles de X_2.

En outre, si la loi du couple (X_1, X_2) détermine de manière unique les lois marginales de X_1 et de X_2, la réciproque n'est pas vraie comme nous le discutons ci-dessous via le concept d'indépendance.

1.4.2 Indépendance

Considérer des familles de variables aléatoires permet en outre de formaliser un concept essentiel de la théorie des probabilités, à savoir la notion d'**indépendance**.

Définition 1.4.2. Soit J un ensemble dénombrable : pour chaque $j \in J$ on considère $A_j \in \mathcal{F}$ un évènement et $X_j : (\Omega, \mathcal{F}) \to (\Omega_j, \mathcal{F}_j)$ une variable aléatoire.
Les évènements $(A_j, j \in J)$ sont **indépendants** si pour tout ensemble fini $I \subset J$ on a

$$\mathbb{P}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbb{P}\left(A_i\right).$$

Les variables aléatoires $(X_j, j \in J)$ sont **indépendantes** si pour toute famille $(A'_j, j \in J)$ d'évènements $A'_j \in \mathcal{F}'$, les évènements $(X_j^{-1}(A'_j), j \in J)$ sont indépendants.
Si $J = \mathbb{N}$, on dit que les variables aléatoires $(X_n, n \in \mathbb{N})$ sont **indépendantes et identiquement distribuées**, en abrégé **i.i.d.**, si elles sont indépendantes et de même loi, i.e., $\mathbb{P}_{X_n} = \mathbb{P}_{X_0}$ pour tout $n \in \mathbb{N}$.

En particulier, deux variables aléatoires à valeurs réelles X_1 et X_2 sont indépendantes si pour tout $A, B \subset \mathbb{R}$, on a

$$\mathbb{P}(X_1 \in A, X_2 \in B) = \mathbb{P}(X_1 \in A)\mathbb{P}(X_2 \in B).$$

Théorème 1.4.2. *Les variables aléatoires X_1 et X_2 sont indépendantes si et seulement si la loi du couple (X_1, X_2) est égale au produit des lois marginales :*

$$\mathbb{P}_{(X_1,X_2)}(\{(x_1,x_2)\}) = \mathbb{P}_{X_1}(\{x_1\})\mathbb{P}_{X_2}(\{x_2\}),\ x_1 \in X_1(\Omega), x_2 \in X_2(\Omega).$$

Dans le cas de l'exemple 1.1.6 où l'on jette deux dés discernables, on vérifie bien que cette définition d'indépendance implique que les variables aléatoires X_1 et X_2 qui enregistrent les résultats de chacun des dés sont indépendantes. Une vision plus intuitive de l'indépendance correspond à dire que

X_1 et X_2 sont indépendantes si toute information sur X_1 ne change en rien la loi de X_2.

En d'autres termes, on a une expérience aléatoire avec un couple (X_1, X_2). Si par exemple X_1 est le minimum des deux dés et X_2 le maximum, la connaissance de X_1 influence la loi de X_2 : bien qu'en général on peut avoir $X_2 = 1$, si l'on sait que $X_1 = 3$ alors cela contraint $X_2 \geq 3$. Si en revanche X_1 est le résultat du premier dé et X_2 du second, alors ces variables aléatoires sont bien indépendantes : le résultat d'un dé n'apporte aucune information sur le résultat de l'autre. La notion de conditionnement introduite dans les Sections 1.6 et 1.7 permettra de formaliser cette idée de "connaissance" ou d'"information" apportée par un évènement et plus généralement par une variable aléatoire. Le résultat suivant est très naturel : si deux variables aléatoires sont indépendantes, alors toute transformation de ces variables aléatoires conserve cette indépendance.

Proposition 1.4.3. *Si $X_1 : (\Omega, \mathcal{F}) \to (\Omega_1, \mathcal{F}_1)$ et $X_2 : (\Omega, \mathcal{F}) \to (\Omega_2, \mathcal{F}_2)$ sont indépendantes, alors pour toutes fonctions $\varphi_1 : (\Omega_1, \mathcal{F}_1) \to (\Omega'_1, \mathcal{F}'_1)$ et $\varphi_2 : (\Omega_2, \mathcal{F}_2) \to (\Omega'_2, \mathcal{F}'_2)$ les variables aléatoires $\varphi_1(X_1)$ et $\varphi_2(X_2)$ sont indépendantes.*

On conclut enfin par deux formules utiles dans le cas de variables aléatoires indépendantes.

Proposition 1.4.4. *Soit X_1 et X_2 deux variables aléatoires indépendantes à valeurs réelles.*

- *Si X_1 et X_2 sont intégrables, alors X_1X_2 est intégrable et $\mathbb{E}(X_1X_2) = \mathbb{E}(X_1)\mathbb{E}(X_2)$;*
- *Si X_1 et X_2 sont de carré intégrables, alors pour tout $a, b \in \mathbb{R}$, $aX_1 + bX_2$ est de carré intégrable et $\mathbb{V}\mathrm{ar}(aX_1 + bX_2) = a^2\mathbb{V}\mathrm{ar}(X_1) + b^2\mathbb{V}\mathrm{ar}(X_2)$.*

On généralise immédiatement la dernière relation : si les X_n à valeurs réelles sont indépendantes et de carré intégrable, alors pour tout $a_1, \ldots, a_n \in \mathbb{R}$ on a

$$\mathbb{V}\mathrm{ar}\left(\sum_{k=1}^{n} a_k X_k\right) = \sum_{k=1}^{n} a_k^2 \mathbb{V}\mathrm{ar}\left(X_k\right).$$

1.4.3 Lois marginales et loi jointe

Nous pouvons maintenant finir la discussion initiée à la fin de la Section 1.4.1 et présenter des exemples de couples de variables aléatoires avec des lois différentes mais dont les lois marginales sont les mêmes. L'idée de base est la suivante :

Les lois marginales de X_1 et X_2 ne contiennent aucune information sur la corrélation entre X_1 et X_2.

Ainsi, prenons deux exemples extrêmes où X_1 et X_2 sont indépendantes et X_1 et X_2 sont parfaitement corrélées : dans les deux exemples, les lois marginales sont les mêmes mais la loi jointe du couple est très différente.

Indépendance : On lance un dé non biaisé deux fois de suite : X_1 enregistre le résultat du premier lancer et X_2 du deuxième. La loi jointe du couple (X_1, X_2) est la loi uniforme sur $\{1, \dots, 6\} \times \{1, \dots, 6\}$ et la loi marginale de X_1 et X_2 est la loi uniforme sur $\{1, \dots, 6\}$;

Corrélation parfaite : On lance un dé : X_1 enregistre le résultat du lancer et pour le "deuxième" lancer on copie le résultat du premier lancer, i.e., on prend $X_2 = X_1$. Alors la loi jointe du couple (X_1, X_2) est la loi uniforme sur $\{(1,1), \dots, (6,6)\}$ alors que les lois marginales de X_1 et X_2 restent inchangées.

1.4.4 Covariance de deux variables aléatoires réelles

Soient X_1, X_2 à valeurs réelles et chacune de carré intégrable. La covariance entre X_1 et X_2 est une certaine mesure de la dépendance, ou plus précisément de la corrélation, entre ces deux variables.

Lemme 1.4.5. *Si X_1 et X_2 à valeurs réelles sont de carré intégrable, alors la variable aléatoire $(X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))$ est intégrable.*

Démonstration. L'inégalité de Cauchy–Schwarz (Théorème 1.3.13) donne

$$\mathbb{E}\left(|X_1 - \mathbb{E}(X_1)|\,|X_2 - \mathbb{E}(X_2)|\right) \leq \sigma_{X_1}\sigma_{X_2}$$

ce qui prouve le résultat. ■

Définition 1.4.3. Soient X_1 et X_2 à valeurs réelles et de carré intégrable : la **covariance entre X_1 et X_2**, notée $\mathbb{C}\mathrm{ov}(X_1, X_2)$, est le nombre réel défini par

$$\mathbb{C}\mathrm{ov}(X_1, X_2) = \mathbb{E}\left[(X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))\right] \in \mathbb{R}.$$

Le nombre $\mathbb{E}[(X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))]/(\sigma_X\sigma_Y) \in [-1, 1]$ est appelé **coefficient de corrélation entre X_1 et X_2**.

Le fait que $\mathbb{E}[(X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))]/(\sigma_X\sigma_Y) \in [-1, 1]$ découle de l'inégalité de Cauchy–Schwarz, cf. preuve du Lemme 1.4.5. On notera en outre que $\mathbb{C}\mathrm{ov}(X_1, X_1) = \mathbb{V}\mathrm{ar}(X_1)$ et que, pour X_1 et X_2 à valeurs dans $\mathbb{R}$, $\mathbb{C}\mathrm{ov}(X_1, X_2) = \mathbb{C}\mathrm{ov}(X_2, X_1)$.

Proposition 1.4.6. *Si X_1 et X_2 à valeurs réelles sont indépendantes et intégrables, alors $\mathbb{C}\mathrm{ov}(X_1, X_2)$ est bien définie et $\mathbb{C}\mathrm{ov}(X_1, X_2) = 0$.*

Démonstration. Si X_1 et X_2 sont indépendantes et intégrables, alors les variables aléatoires $X_1 - \mathbb{E}(X_1)$ et $X_2 - \mathbb{E}(X_2)$ le sont aussi et donc le fait que $\mathbb{C}\mathrm{ov}(X_1, X_2)$ existe et soit nul découle de la Proposition 1.4.4. ■

Attention :

La réciproque n'est en général pas vraie,
i.e., covariance nulle N'IMPLIQUE PAS (en général) indépendance.

Un contre-exemple simple est obtenu en considérant $X_1 = \pm 1$ avec probabilité $1/2$ et $X_2 = 0$ si $X_1 = 1$ et $X_2 = \pm 1$ avec probabilité $1/2$ sinon. Alors $\mathbb{E}(X_1) = \mathbb{E}(X_2) = \mathbb{E}(X_1 X_2) = 0$ bien que X_1 et X_2 ne soient pas indépendantes.

Si covariance nulle n'implique en général pas indépendance, il existe un cas particulier très important où cela est vrai : il s'agit du cas des vecteurs gaussiens traités dans le Chapitre 4.

1.4.5 Espérance, variance et covariance : généralisation au cas vectoriel et matriciel

On généralise maintenant les notions d'espérance, de variance et de covariance au cas vectoriel. Le formalisme introduit est très puissant et permet notamment de traiter de manière unifier les variables aléatoires à valeurs dans $\mathbb{R}^d$ avec $d \geq 1$.

Définition 1.4.4. Soit X une variable aléatoire discrète à valeurs dans $\mathbb{R}^{m\times n}$, i.e., $X = (X_{ij})_{1\leq i\leq m, 1\leq j\leq n}$ avec chaque X_{ij} à valeurs dans $\mathbb{R}$.
On dit que X est **intégrable** si chaque X_{ij} l'est, et **de carré intégrable** si chaque X_{ij} l'est.

Définition 1.4.5. Si $X = (X_{ij})_{1\leq i\leq m, 1\leq j\leq n}$ à valeurs dans $\mathbb{R}^{m\times n}$ est intégrable, on définit $\mathbb{E}(X) \in \mathbb{R}^{m\times n}$ comme la matrice dont chaque entrée est la moyenne de l'entrée correspondante de X : $\mathbb{E}(X) = (\mathbb{E}(X_{ij}))_{1\leq i\leq m, 1\leq j\leq n}$.

L'opérateur $\mathbb{E}$ ainsi défini reste linéaire, et satisfait en outre $(\mathbb{E}(X))^T = \mathbb{E}(X^T)$. On rappelle que dans le cas scalaire, covariance et variance sont définies par $\mathbb{C}\mathrm{ov}(X_1, X_2) = \mathbb{E}[(X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))]$ et $\mathbb{V}\mathrm{ar}(X) = \mathbb{C}\mathrm{ov}(X, X) = \mathbb{E}[(X - \mathbb{E}(X))^2]$: les définitions ci-dessous permettent de généraliser ces définitions au cas vectoriel.

Définition 1.4.6. Soit X_1 et X_2 à valeurs dans $\mathbb{R}^{n_1}$ et $\mathbb{R}^{n_2}$, respectivement, supposées de carré intégrable. La **(matrice de) covariance entre X_1 et X_2**, notée $\mathbb{C}\mathrm{ov}(X_1, X_2)$, est la matrice de taille $n_1 \times n_2$ à valeurs dans $\mathbb{R}$ définie de la manière suivante :

$$\mathbb{C}\mathrm{ov}(X_1, X_2) = \mathbb{E}\left[(X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))^T\right] \in \mathbb{R}^{n_1\times n_2}.$$

Du fait que $(\mathbb{E}(X))^T = \mathbb{E}(X^T)$, il s'ensuit que $\mathbb{C}\mathrm{ov}(X_1, X_2) = \mathbb{C}\mathrm{ov}(X_2, X_1)^T$.

Définition 1.4.7. Soit X à valeurs dans $\mathbb{R}^n$ de carré intégrable. La **matrice de covariance de X**, notée $\mathbb{V}\mathrm{ar}(X)$, est la matrice symétrique carrée de taille $n \times n$ définie par :

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{C}\mathrm{ov}(X, X) = \mathbb{E}\left[(X - \mathbb{E}(X))(X - \mathbb{E}(X))^T\right] \tag{1.9}$$

ou de manière plus explicite, $\mathbb{V}\mathrm{ar}(X) = (\mathbb{C}\mathrm{ov}(X_i, X_j), 1 \leq i, j \leq n)$:

$$\mathbb{V}\mathrm{ar}(X) = \begin{pmatrix} \mathbb{V}\mathrm{ar}(X_1) & \mathbb{C}\mathrm{ov}(X_1, X_2) & \cdots & \mathbb{C}\mathrm{ov}(X_1, X_n) \\ \mathbb{C}\mathrm{ov}(X_2, X_1) & \mathbb{V}\mathrm{ar}(X_2) & \cdots & \mathbb{C}\mathrm{ov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}\mathrm{ov}(X_n, X_1) & \mathbb{C}\mathrm{ov}(X_n, X_2) & \cdots & \mathbb{V}\mathrm{ar}(X_n) \end{pmatrix}. \tag{1.10}$$

La proposition suivante est fondamentale dans l'étude des vecteurs gaussiens.

Proposition 1.4.7. *La matrice* $\mathbb{V}\mathrm{ar}(X)$ *est symétrique et positive.*

Démonstration. La symétrie vient du fait que $\mathbb{C}\mathrm{ov}(X_1, X_2)^T = \mathbb{C}\mathrm{ov}(X_2, X_1)$, et le caractère positif du fait que pour tout $x \in \mathbb{R}^n$, si

$$x^T \mathbb{V}\mathrm{ar}(X) x = x^T \mathbb{E}(YY^T) x = \mathbb{E}\left((x^T Y)^2\right)$$

où l'on a noté $Y = X - \mathbb{E}(X)$ pour alléger les notations. ■

L'avantage de ces notations vectorielles et matricielles est qu'elles permettent de manière transparente des manipulations "par bloc". En effet, si on a la décomposition par bloc

$$X = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nn} \\ \vdots & \vdots & \vdots & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mn} \end{pmatrix}$$

où les X_{ij} sont elles-mêmes des matrices aléatoires intégrables, alors on a la même décomposition par bloc pour $\mathbb{E}(X)$:

$$\mathbb{E}(X) = \begin{pmatrix} \mathbb{E}(X_{11}) & \mathbb{E}(X_{12}) & \cdots & \mathbb{E}(X_{1n}) \\ \mathbb{E}(X_{21}) & \mathbb{E}(X_{22}) & \cdots & \mathbb{E}(X_{2n}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}(X_{m1}) & \mathbb{E}(X_{m2}) & \cdots & \mathbb{E}(X_{mn}) \\ \mathbb{E}(X_{n1}) & \mathbb{E}(X_{n2}) & \cdots & \mathbb{E}(X_{nn}) \\ \vdots & \vdots & \vdots & \vdots \\ \mathbb{E}(X_{m1}) & \mathbb{E}(X_{m2}) & \cdots & \mathbb{E}(X_{mn}) \end{pmatrix}.$$

De la même manière, si on la décomposition par bloc

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$

avec chaque X_k un vecteur aléatoire de carré intégrable, alors la formule (1.10) reste valable.

1.4.6 Remarques

Dans la présentation ci-dessus, nous sommes partis d'une famille de variables aléatoires $(X_n, n \in \mathbb{N})$ définies sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ et avons considéré les différentes marginales X_1, X_2, etc. Nous concluons cette section par deux remarques.

Tout d'abord, la construction inverse est possible : si l'on part de deux variables aléatoires $X_1 : (\Omega_1, \mathcal{F}_1, \mathbb{P}_1) \to (\Omega_1', \mathcal{F}_1')$ et $X_2 : (\Omega_2, \mathcal{F}_2, \mathbb{P}_2) \to (\Omega_2', \mathcal{F}_2')$, alors on peut définir la variable aléatoire $X = (X_1, X_2)$ sur l'espace produit $\Omega = \Omega_1 \times \Omega_2$ muni de la tribu $\mathcal{F} = \mathcal{P}(\Omega)$. Considérer des variables aléatoires indépendantes revient alors à considérer la mesure tensorielle $\mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_2$ sur Ω définie par $\mathbb{P}(A_1 \times A_2) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2)$ pour $A_1 \in \mathcal{F}_1$ et $A_2 \in \mathcal{F}_2$, mais d'autres mesures sur l'espace produit sont possibles, reflétant toutes les

relations de dépendance possibles entre deux variables aléatoires X_1 et X_2 dont seulement les marginales sont spécifiées.

En outre, cette approche présente un trou théorique : si X_n est une variable aléatoire sur $(\Omega, \mathcal{F}, \mathbb{P})$, alors considérer une suite $(X_n, n \in \mathbb{N})$ de variables i.i.d. revient à considérer la variable aléatoire $X = (X_n, n \in \mathbb{N})$ sur l'espace produit $\Omega^{\mathbb{N}}$ muni de la mesure tensorielle $\mathbb{P}^{\otimes\mathbb{N}}$. Néanmoins, $\Omega^{\mathbb{N}}$ n'est plus dénombrable et l'on sort donc du cadre discret de ce chapitre ! Plus généralement, si l'on veut considérer une suite infinies de variables aléatoires indépendantes, alors l'espace de probabilité sous-jacent doit être suffisamment gros et en particulier, non-dénombrable. Pour pallier ce problème, nous aurions pu faire le choix de nous restreindre à des suites finies dans tout ce chapitre, mais considérer des suites infinies est indispensable pour les théorèmes limites du Chapitre 3 et nous nous accommoderons donc de ce trou théorique.

1.5 Fonction caractéristique, fonction génératrice et transformée de Laplace

1.5.1 Le cas de la dimension un

On définit plusieurs transformées liées à une variable aléatoire discrète à valeurs réelles. Pour une variable aléatoire Z à valeurs complexes, on définit son espérance par $\mathbb{E}(Z) = \mathbb{E}(\Re(Z)) + i\mathbb{E}(\Im(Z))$ dès lors que $\Re(Z)$ et $\Im(Z)$, les parties réelle et imaginaire de Z, sont intégrables.

Définition 1.5.1. Soit X à valeurs dans $\mathbb{R}$: sa **fonction caractéristique** est la fonction $\phi_X : \mathbb{R} \to \mathbb{C}$ définie de la manière suivante :

$$\varphi_X(t) = \mathbb{E}\left(e^{itX}\right), \ t \in \mathbb{R}.$$

Définition 1.5.2. Soit X à valeurs dans $\mathbb{R}_+$: sa **transformée de Laplace** est la fonction $L_X : \mathbb{R}_+ \to \mathbb{R}_+$ définie de la manière suivante :

$$L_X(\lambda) = \mathbb{E}\left(e^{-\lambda X}\right), \ \lambda \in \mathbb{R}_+.$$

Définition 1.5.3. Soit X à valeurs dans $\mathbb{N}$: sa **fonction génératrice** est la fonction $\phi_X : [-1, 1] \to [-1, 1]$ définie par

$$\phi_X(z) = \mathbb{E}\left(z^X\right), \ z \in [-1, 1].$$

Proposition 1.5.1. *Chacune de ces transformées (fonction caractéristique, transformée de Laplace ou fonction génératrice) caractérise la loi de X. Par exemple, si X et Y sont à valeurs dans $\mathbb{R}$ et que $\phi_X = \phi_Y$, alors $\mathbb{P}_X = \mathbb{P}_Y$.*

Les transformées associées aux lois classiques introduites en Section 1.2.4 sont données dans le Tableau B.1 en page 174.

1.5.2 Le cas multi-dimensionnel

Les définitions ci-dessus s'étendent naturellement au cas multi-dimensionnel : pour la fonction caractéristique et la transformée de Laplace, il suffit de remplacer le produit par un produit scalaire :

$$\varphi_X(t) = \mathbb{E}\left(e^{i\langle t, X\rangle}\right), \ t \in \mathbb{R}^n, \quad \text{et} \quad L_X(\lambda) = \mathbb{E}\left(e^{-\langle \lambda, X\rangle}\right), \ \lambda \in \mathbb{R}_+^n.$$

Pour la fonction génératrice, on introduit la notation $z^x = z_1^{x_1} \cdots z_n^{x_n}$ pour $z = (z_1, \ldots, z_n) \in \mathbb{R}^n$ et $x = (x_1, \ldots, x_n) \in \mathbb{R}_+^n$, si bien que

$$\phi_X(z) = \mathbb{E}\left(z^X\right), \ z \in [-1, 1]^n.$$

La Proposition 1.5.1 reste vraie, i.e., chacune de ces transformées caractérise la loi de X. En outre, ces transformées peuvent aussi être utiles pour caractériser l'indépendance.

Proposition 1.5.2. *Soit X_k à valeurs dans $\mathbb{R}^{n_k}$: alors $(X_1, \ldots, X_n)$ sont indépendantes si et seulement si*

$$\varphi_X(t) = \prod_{k=1}^{n} \varphi_{X_k}(t_k), \ t = (t_1, \ldots, t_n) \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_n}.$$

Si les X_k sont à valeurs dans $\mathbb{R}_+^{n_k}$, alors $(X_1, \ldots, X_n)$ sont indépendantes si et seulement si

$$L_X(\lambda) = \prod_{k=1}^{n} L_{X_k}(\lambda_k), \ \lambda = (\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_n}.$$

Si les X_k sont à valeurs dans $\mathbb{N}^{n_k}$, alors $(X_1, \ldots, X_n)$ sont indépendantes si et seulement si

$$\phi_X(z) = \prod_{k=1}^{n} \phi_{X_k}(z_k), \ z = (z_1, \ldots, z_n) \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_n}.$$

1.5.3 Lien avec l'analyse harmonique

Les trois transformées introduites ci-dessus correspondent aux transformées de Fourier, transformée en z (Tz) et transformée de Laplace (TL) introduites dans le chapitre VI d'analyse harmonique du polycopié de mathématiques déterministes. On notera néanmoins des différences de notation et de convention, par exemple :

- La fonction caractéristique $\varphi_X(t) = \mathbb{E}(e^{itX})$ peut se réécrire

  $$\varphi_X(t) = \sum_{x \in X(\Omega)} \mathbb{P}_X(\{x\}) e^{itx}$$

  et correspond donc, pour $X(\Omega) = \mathbb{Z}$ et à un facteur -2π près, à la transformée de Fourier d'une suite apériodique (TFed) pour une suite dans $\ell^1(\mathbb{Z})$, cf. Définition 11.2.1 du polycopié de mathématiques déterministes. En outre, si cette transformée TFed a une interprétation physique entre domaine temporel et fréquentiel (ce qui justifie le facteur -2π) la fonction caractéristique n'a pas cette interprétation ;
- La fonction génératrice $\phi_X(z) = \mathbb{E}(z^X)$ peut se réécrire

  $$\phi_X(z) = \sum_{n \in \mathbb{N}} \mathbb{P}_X(\{n\}) z^n$$

  et correspond donc à la transformée en z (Tz) de la section 12.1 du polycopié de mathématiques déterministes, à la différence que la Tz est définie comme

  $$\sum_{n \in \mathbb{N}} \mathbb{P}_X(\{n\}) z^{-n}$$

  pour $z \in \mathbb{C}$ tel que la série soit convergente.

Enfin, le lien avec la transformée de Laplace de la section 12.2 du polycopié de mathématiques déterministes sera plus explicite au chapitre suivant lorsque nous aurons introduit les variables aléatoires absolument continues. Dans ce cadre, la fonction caractéristique devient alors la transformée de Fourier d'une fonction apériodique (TFec) de la définition 11.4.1 du polycopié de mathématiques déterministes.

1.6 Conditionnement par rapport à un évènement

1.6.1 Définition

Les deux dernières parties du chapitre sont dédiées à la notion de conditionnement : on commence ici par le **conditionnement par rapport à un évènement**, qui s'inscrit dans la continuité du programme de classes préparatoires, et on généralise cette notion dans la section suivante en introduisant le **conditionnement par rapport à une variable aléatoire**.

Définition 1.6.1. Soient $A, B \in \mathcal{F}$. La **probabilité conditionnelle de B sachant A**, notée $\mathbb{P}(B \mid A)$, est définie par

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

avec la convention $\mathbb{P}(B \mid A) = 0$ si $\mathbb{P}(A) = 0$.

Théorème 1.6.1. *Si $\mathbb{P}(A) > 0$, la fonction $\mathbb{P}(\cdot \mid A) : B \in \mathcal{F} \mapsto \mathbb{P}(B \mid A)$ est une mesure de probabilité sur $(\Omega, \mathcal{F})$ appelée **mesure de probabilité conditionnelle sachant** A.*

Ainsi, tout évènement $A \in \mathcal{F}$ tel que $\mathbb{P}(A) > 0$ induit une **nouvelle mesure de probabilité** sur $(\Omega, \mathcal{F})$, à savoir la probabilité conditionnelle $\mathbb{P}(\cdot \mid A)$. Comme souligné dans la Remarque 1.3.1, à chaque mesure de probabilité sur $(\Omega, \mathcal{F})$ correspond une espérance, ce qui nous amène naturellement à définir l'espérance conditionnelle par rapport à un évènement comme l'espérance associée à la mesure de probabilité $\mathbb{P}(\cdot \mid A)$.

Définition 1.6.2. Soit $A \in \mathcal{F}$ avec $\mathbb{P}(A) > 0$. L'espérance conditionnelle sachant A, notée $\mathbb{E}(\cdot \mid A)$, est l'espérance associée à la mesure de probabilité $\mathbb{P}(\cdot \mid A)$. Pour A avec $\mathbb{P}(A) = 0$ on adoptera la convention $\mathbb{E}(X \mid A) = 0$ pour toute variable aléatoire X.

Par la suite, on utilisera la notation $\mathbb{E}(X\xi_A) = \mathbb{E}(X; A)$ dès lors que l'espérance du membre de gauche est bien définie, par exemple si $X \geq 0$ (on remarquera que si X est discrète, alors $X\xi_A$ l'est aussi).

Théorème 1.6.2. *Soit $A \in \mathcal{F}$ avec $\mathbb{P}(A) > 0$ et X une variable aléatoire discrète à valeurs réelles. Si $X \geq 0$, alors*

$$\mathbb{E}(X \mid A) = \frac{\mathbb{E}(X\xi_A)}{\mathbb{P}(A)} = \frac{\mathbb{E}(X; A)}{\mathbb{P}(A)}.$$

Cette formule reste valable si X discrète à valeurs réelles est intégrable.

Démonstration. Puisque $\mathbb{E}(\cdot \mid A)$ est l'opérateur d'espérance associé à la mesure de probabilité $\mathbb{P}(\cdot \mid A)$, pour $X \in \mathbb{R}_+$ on a par définition de l'espérance

$$\mathbb{E}(X \mid A) = \sum_{x \in X(\Omega)} x\mathbb{P}(X = x \mid A)$$

puis par définition de $\mathbb{P}(\cdot \mid A)$,

$$\mathbb{E}(X \mid A) = \frac{1}{\mathbb{P}(A)} \sum_{x \in X(\Omega)} x\mathbb{P}(\{X = x\} \cap A).$$

Pour tout $x \in \mathbb{R}$, on vérifie que $x\mathbb{P}(\{X = x\} \cap A) = x\mathbb{P}(\xi_A X = x)$: cela donne le résultat pour $X \geq 0$ et le résultat pour $X \in \mathbb{R}$ intégrable s'ensuit en revenant aux parties positive et négative. ■

1.6.2 Retour sur l'indépendance

On revient sur l'encart de la Section 1.4.2 : la notion de conditionnement permet de reformuler le fait que deux variables aléatoires X_1 et X_2 sont indépendantes si toute information sur X_1 ne change pas la loi de X_2.

Théorème 1.6.3. *Soient A et B deux évènements : alors*

$$A \text{ et } B \text{ sont indépendants} \iff \mathbb{P}(B \mid A) = \mathbb{P}(B) \iff \mathbb{P}(A \mid B) = \mathbb{P}(A).$$

En outre, $X_1 : (\Omega, \mathcal{F}) \to (\Omega_1, \mathcal{F}_1)$ et $X_2 : (\Omega, \mathcal{F}) \to (\Omega_2, \mathcal{F}_2)$ sont indépendantes si et seulement si

$$\forall A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2,\ \mathbb{P}(X_1 \in A_1 \mid X_2 \in A_2) = \mathbb{P}(X_1 \in A_1). \tag{1.11}$$

1.7 Conditionnement par rapport à une variable aléatoire discrète

1.7.1 Définition de l'espérance conditionnelle par rapport à une variable aléatoire discrète

Les notions introduites ci-dessus constituent un rappel de ce que vous avez vu en classes préparatoires. Nous introduisons maintenant une nouvelle notion : l'**espérance conditionnelle par rapport à une variable aléatoire**. On commence par donner la définition de l'espérance conditionnelle par rapport à une variable aléatoire, que l'on discute par la suite.

Définition 1.7.1. Soit $X : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$ une variable aléatoire discrète quelconque et $Y : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{F})$ une variable aléatoire discrète à valeurs réelles et intégrable. L'espérance conditionnelle de Y par rapport à X, dénotée $\mathbb{E}(Y \mid X)$, est la **variable aléatoire** $h(X)$ où $h : (\Omega', \mathcal{F}') \to (\mathbb{R}, \mathcal{P}(\mathbb{R}))$ est la fonction définie par

$$h(x) = \mathbb{E}(Y \mid X = x),\ x \in X(\Omega).$$

On retiendra notamment de cette définition que

$\mathbb{E}(Y \mid X)$ est une variable aléatoire ! !

Puisque $\mathbb{E}(Y \mid X) = h(X)$, $\mathbb{E}(Y \mid X)$ est même une variable aléatoire entièrement déterminée par X.

Dans toute cette section, $X : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$ est une variable aléatoire discrète quelconque et $Y : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{F})$ est une variable aléatoire discrète à valeurs réelles et intégrable.

1.7.2 Intuition

Intuitivement, l'espérance conditionnelle de Y sachant X peut être définie de la manière suivante :

L'espérance conditionnelle $\mathbb{E}(Y \mid X)$ correspond à moyenniser tout ce qui détermine Y mais qui n'est pas déterminé par X.

Essayons d'éclairer notre propos. Pour tout couple (X, Y), X explique en partie plus ou moins grande la valeur prise par Y. Par exemple, X peut complètement déterminer Y si $Y = \varphi(X)$ est une fonction de X, ou bien au contraire X et Y peuvent être indépendantes auquel cas X n'a aucun pouvoir prédictif sur Y.

Dans le cas général, imaginons que l'on puisse écrire $Y = \varphi(X, Z)$ avec X et Z indépendantes, où Z correspond intuitivement à ce qui détermine la valeur de Y mais ne peut pas être expliqué par X (d'où l'indépendance entre X et Z). Dans l'encart ci-dessus, "moyenniser tout ce qui détermine Y mais qui n'est pas déterminé par X" signifie alors précisément moyenniser sur Z et pas sur X, ce qui suggère naturellement de définir

$$\mathbb{E}(Y \mid X) = \sum_{z \in Z(\Omega)} \varphi(X, z)\mathbb{P}(Z = z) \tag{1.12}$$

qui correspond bien à la moyenne de $\varphi(X, Z)$ en gardant X constant. On remarquera que le membre de droite est une fonction de X, disons $h(X)$, en particulier c'est une variable aléatoire et non un nombre déterministe. En effet, on insiste sur le fait que

$\mathbb{E}(Y \mid X)$ est une variable aléatoire ! !

Par ailleurs, la fonction h telle que (1.12) se réécrit $\mathbb{E}(Y \mid X) = h(X)$ est la fonction

$$h(x) = \sum_{z \in Z(\Omega)} \varphi(x, z)\mathbb{P}(Z = z)$$

et l'on vérifie bien, en utilisant l'indépendance de X et Z, que $h(x) = \mathbb{E}(Y \mid X = x)$ en conformité avec la définition générale. Par ailleurs, cette définition intuitive permet de calculer directement $\mathbb{E}(Y \mid X)$ dans les deux cas extrêmes où X et Y sont parfaitement corrélées ou au contraire indépendantes :

- si $Y = \varphi(X)$, alors on n'a pas besoin de la variable explicative Z et donc
$$\mathbb{E}(Y \mid X) = \varphi(X) = Y;$$
- si au contraire X et Y sont indépendantes, alors on n'a pas besoin de X et donc
$$\mathbb{E}(Y \mid X) = \sum_{z \in Z(\Omega)} \varphi(z)\mathbb{P}(Z = z) = \mathbb{E}(\varphi(Z)) = \mathbb{E}(Y).$$

1.7.3 Propriétés

On commence par prouver que les résultats "intuitifs" précédents sont effectivement corrects.

Proposition 1.7.1. *Si $Y = \varphi(X)$, alors $\mathbb{E}(Y \mid X) = Y$.*
Si Y est indépendante de X alors $\mathbb{E}(Y \mid X) = \mathbb{E}(Y)$.
Plus généralement, si $Y = \varphi(X, Z)$ avec Z une variable aléatoire discrète indépendante de X et que pour chaque $x \in X(\Omega)$ la variable aléatoire $\varphi(x, Z)$ est intégrable, alors la formule (1.12) est valide, i.e.,

$$\mathbb{E}(Y \mid X) = \sum_{z \in Z(\Omega)} \varphi(X, z)\mathbb{P}_Z(\{z\}).$$

Démonstration. Si X et Y sont indépendantes, alors $h(x) = \mathbb{E}(Y \mid X = x) = \mathbb{E}(Y)$.
Si $Y = \varphi(X)$, alors $h(x) = \mathbb{E}(\varphi(X) \mid X = x) = \mathbb{E}(\varphi(x) \mid X = x) = \varphi(x)$ et donc $\mathbb{E}(Y \mid X) = h(X) = \varphi(X) = Y$.
Si $Y = \varphi(X, Z)$ avec X et Z indépendantes, alors

$$h(x) = \mathbb{E}(\varphi(X, Z) \mid X = x) = \mathbb{E}(\varphi(x, Z) \mid X = x) = \mathbb{E}(\varphi(x, Z)) = \sum_{z \in Z(\Omega)} \varphi(x, z)\mathbb{P}_Z(\{z\})$$

qui donne le résultat. ■

Par ailleurs, l'espérance conditionnelle satisfait toutes les propriétés de l'espérance classique.

Proposition 1.7.2. *Soit Y_1, Y_2 des variables aléatoires discrètes réelles et intégrables.*

- *Si $Y_2 \geq Y_1$, alors $\mathbb{E}(Y_2 \mid X) \geq \mathbb{E}(Y_1 \mid X)$, en particulier, $\mathbb{E}(Y \mid X) \geq 0$ si $Y \geq 0$.*
- *Pour tout $a, b \in \mathbb{R}$, $\mathbb{E}(aY_1 + bY_2 \mid X) = a\mathbb{E}(Y_1 \mid X) + b\mathbb{E}(Y_2 \mid X)$.*
- *Pour toute fonction $\varphi : \Omega' \to \mathbb{R}$, $\mathbb{E}(\varphi(X)Y \mid X) = \varphi(X)\mathbb{E}(Y \mid X)$.*
- *$|\mathbb{E}(Y \mid X)| \leq \mathbb{E}(|Y| \mid X)$.*

Comme d'habitude, pour des variables aléatoires positives toutes les relations ci-dessus restent valables même si les variables aléatoires en jeu ne sont pas intégrables. L'espérance de l'espérance conditionnelle satisfait une propriété bien particulière.

Théorème 1.7.3 (Théorème de l'espérance totale)**.** *La variable aléatoire $\mathbb{E}(Y \mid X)$ est intégrable et son espérance est donnée par*

$$\mathbb{E}\big(\mathbb{E}(Y \mid X)\big) = \mathbb{E}(Y).$$

Démonstration. On traite le cas où $Y \geq 0$, le cas général est obtenu comme d'habitude en décomposant en parties positive et négative. Dans ce cas, la Proposition 1.7.2 implique que $\mathbb{E}(Y \mid X) = h(X) \geq 0$ et donc le Théorème 1.3.6 donne

$$\mathbb{E}\big(\mathbb{E}(Y \mid X)\big) = \sum_{x \in X(\Omega)} h(x)\mathbb{P}(X = x).$$

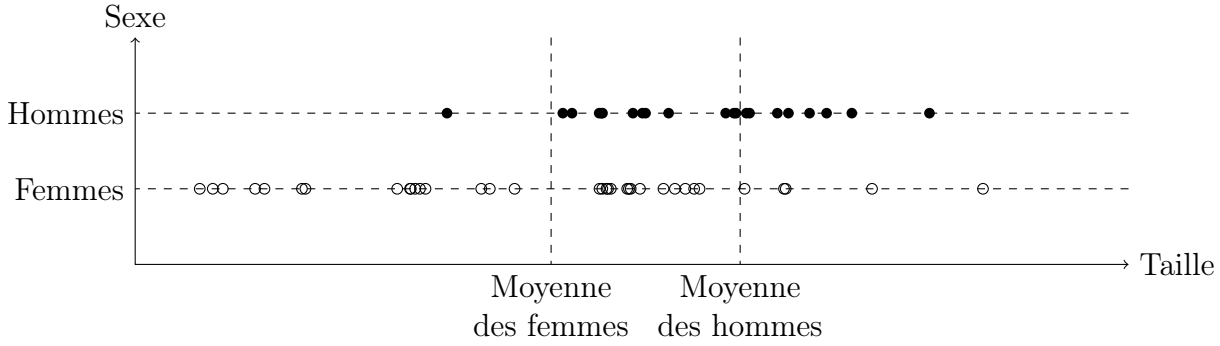


FIGURE 1.1 – Chaque point représente la taille d'un individu et les points sont regroupés par sexe. Pour obtenir la moyenne des tailles, le théorème de l'espérance totale nous dit qu'on peut faire la moyenne par groupes, puis faire la moyenne des moyennes en tenant compte de l'importance relative des deux groupes.

Par définition de $h(x)$, on a $h(x)\mathbb{P}(X=x)=0$ pour $x \in X(\Omega)$ avec $\mathbb{P}(X=x)=0$, et pour $x \in X(\Omega)$ avec $\mathbb{P}(X=x)>0$ on a

$$h(x)\mathbb{P}(X=x) = \mathbb{E}(Y \mid X=x)\mathbb{P}(X=x) = \mathbb{E}(Y; X=x) = \mathbb{E}(Y\mathbb{1}\{X=x\}).$$

Le théorème de Fubini donne alors

$$\mathbb{E}\big(\mathbb{E}(Y \mid X)\big) = \mathbb{E}\left(\sum_{x\in X(\Omega)} Y\mathbb{1}\{X=x\}\right) = \mathbb{E}\left(Y\sum_{x\in X(\Omega)} \mathbb{1}\{X=x\}\right) = \mathbb{E}(Y)$$

puisque $\sum_{x\in X(\Omega)} \mathbb{1}\{X=x\} = 1$. ■

Nous présentons deux exemples typiques d'utilisation du théorème de l'espérance totale. Le premier exemple montre que le théorème de l'espérance totale formalise quelque chose de très intuitif : pour calculer une moyenne, on peut partitionner les valeurs à moyenniser, calculer la moyenne dans chaque groupe et faire la moyenne des moyennes en pondérant par la taille de chaque groupe.

Exemple 1.7.1. On considère la taille d'individus dans une population constituée de 21 hommes et 35 femmes, et on souhaite calculer la moyenne. Pour cela, il y a deux manières :

- on moyennise directement les 56 valeurs $t_1, \dots, t_{56}$, soit $\frac{1}{56}\sum_{k=1}^{56} t_k$;
- on calcule la moyenne par sexe, puis on moyennise les moyennes en pondérant par la taille respective des groupes, soit $\frac{21}{56}$ et $\frac{35}{56}$, cf. Figure 1.1. Cela revient à faire

$$\text{Moyenne} = \frac{35}{56} \times \text{Moyenne des femmes} + \frac{21}{56} \times \text{Moyenne des hommes}. \tag{1.13}$$

La deuxième manière correspond exactement au théorème de l'espérance totale. En effet, soit T la variable aléatoire tirée uniformément au hasard dans l'ensemble $\{t_k\}$ des tailles des individus de la population, si bien que l'on veut calculer

$$\mathbb{E}(T) = \frac{1}{56}\sum_{k=1}^{56} t_k.$$

Soit ξ_A la fonction indicatrice de l'évènement $A = \{\text{on tire une femme}\}$: le théorème de l'espérance totale nous donne

$$\mathbb{E}(T) = \mathbb{E}\left[\mathbb{E}(T \mid \xi_A)\right]$$

ce qui nous permet bien de retrouver (1.13) puisque

$$\mathbb{E}(T \mid \xi_A) = \begin{cases} \text{Moyenne des femmes} & \text{si } \xi_A = 1, \\ \text{Moyenne des hommes} & \text{si } \xi_A = 0 \end{cases}$$

et que

$$\mathbb{P}(\xi_A = 1) = \mathbb{P}(\{\text{on tire une femme}\}) = \frac{\#\text{femmes}}{\#\text{individus}} = \frac{35}{56} = 1 - \mathbb{P}(\xi_A = 0).$$

Exemple 1.7.2. Cet exemple montre l'utilité du théorème de l'espérance totale dans un cadre plus théorique, où ce résultat est très utile pour effectuer certains calculs. Soit S la somme d'un nombre aléatoire de variables aléatoires i.i.d. :

$$\mathbb{E}(S) = \sum_{k=1}^{N} X_k$$

où les variables aléatoires $(X_n, n \in \mathbb{N}^*)$ sont i.i.d. et $N \in \mathbb{N}^*$ est indépendante des X_n. On montre aisément à l'aide du Théorème 1.7.3 que $\mathbb{E}(S) = \mathbb{E}(N)\mathbb{E}(X)$. Pour cela, on calcule d'abord $\mathbb{E}(S \mid N)$: on écrit $S = \varphi(X, N)$ avec $\varphi(x, n) = \sum_{k=1}^{n} x_k$, si bien que

$$\mathbb{E}(S \mid N = n) = \mathbb{E}(\varphi(X, n)) = \mathbb{E}\left(\sum_{k=1}^{n} X_k\right) = n\mathbb{E}(X)$$

et qui implique que $\mathbb{E}(S \mid N) = N\mathbb{E}(X)$. On utilise maintenant le théorème de l'espérance totale pour obtenir

$$\mathbb{E}(S) = \mathbb{E}[\mathbb{E}(S \mid N)] = \mathbb{E}(N\mathbb{E}(X)) = \mathbb{E}(N)\mathbb{E}(X).$$

En combinant la Proposition 1.7.2 et le Théorème 1.7.3, on obtient la généralisation suivante du théorème de l'espérance totale.

Proposition 1.7.4. *Pour toute fonction $\varphi : \Omega' \to \mathbb{R}$, on a $\mathbb{E}(\varphi(X)Y) = \mathbb{E}[\varphi(X)\mathbb{E}(Y \mid X)]$.*

Démonstration. Le théorème de l'espérance totale nous assure que

$$\mathbb{E}(\varphi(X)Y) = \mathbb{E}\left[\mathbb{E}(\varphi(X)Y \mid X)\right]$$

et puisque $\mathbb{E}(\varphi(X)Y \mid X) = \varphi(X)\mathbb{E}(Y \mid X)$ par la Proposition 1.7.2 on obtient le résultat. ■

1.7.4 Loi conditionnelle

Dans toutes les sections précédentes, on est partis d'une mesure de probabilité $\mathbb{P}$ ou $\mathbb{P}(\cdot \mid A)$ pour définir l'espérance $\mathbb{E}$ ou $\mathbb{E}(\cdot \mid A)$ associée, et on a montré que probabilités et espérance étaient reliées par $\mathbb{P}(A) = \mathbb{E}(\xi_A)$. Dans le cas de l'espérance conditionnelle par rapport une variable aléatoire, on a directement défini l'espérance, et on utilise la relation précédente pour définir l'espérance conditionnelle associée.

Définition 1.7.2. La **mesure de probabilité conditionnelle sachant** X, notée $\mathbb{P}(\cdot \mid X)$, est la mesure de probabilité aléatoire définie de la manière suivante :

$$\mathbb{P}(A \mid X) = \mathbb{E}\left(\xi_A \mid X\right), \ A \in \mathcal{F}.$$

En particulier, la mesure de probabilité aléatoire $\mathbb{P}_{Y|X}$

$$A \in \mathcal{F} \mapsto \mathbb{P}(Y \in A \mid X)$$

est appelée **loi conditionnelle de Y sachant X**.

Sous réserve d'être bien définies, toutes les relations liant espérance et mesure de probabilité restent vraies conditionnellement à une variable aléatoire. Par exemple, on a le théorème de transfert conditionnel suivant, qui généralise le Théorème 1.3.6.

Théorème 1.7.5. *Soit $\varphi : \Omega' \to \mathbb{R}$: si $\varphi(Y) \geq 0$ ou $\mathbb{E}(|\varphi(Y)|) < \infty$, alors*

$$\mathbb{E}\left(\varphi(Y) \mid X\right) = \sum_{y \in Y(\Omega)} \varphi(y)\mathbb{P}(Y = y \mid X). \tag{1.14}$$

1.7.5 Approche variationnelle et lien avec l'analyse fonctionnelle

Nous présentons maintenant une deuxième manière intuitive de voir l'espérance conditionnelle. Supposons que Y soit de carré intégrable : alors la dérivée de la fonction $y \mapsto \mathbb{E}[(Y-y)^2]$ vaut $2(y - \mathbb{E}(Y))$ et donc l'espérance $\mathbb{E}(Y)$ peut être vue comme la quantité minimisant cette fonction. En d'autres termes, on a que, dans un certain sens (en fait, L_2),

$\mathbb{E}(Y)$ **est la meilleure approximation de Y par une constante.**

La notion d'espérance conditionnelle généralise cette manière de voir l'espérance : c'est en fait même la bonne manière de définir l'espérance conditionnelle dans un contexte général. Généralisant la définition précédente de l'espérance, on peut dire que

$\mathbb{E}(Y \mid X)$ **est la meilleure approximation de Y par une fonction de X**

et formellement, on a en effet, dans le cas où X est de carré intégrable,

$$\mathbb{E}(Y \mid X) = h(X) \quad \text{avec} \quad h = \arg\min \mathbb{E}\left[(Y - g(X))^2\right]$$

où le minimum est pris sur toutes les fonctions mesurables g.

D'un point de vue abstrait, l'espérance conditionnelle n'est donc en fait rien d'autre que la projection orthogonale, dans l'espace de Hilbert des variables aléatoires de carré intégrable, de la variable Y sur l'espace engendré par X.

1.8 Fiche de synthèse

X est une **variable aléatoire discrète** si c'est une application de Ω dans $X(\Omega)$ fini ou dénombrable. L'ensemble (=évènement) $X^{-1}(B) = \{\omega \in \Omega \ ; \ X(\omega) \in B\}$ sera noté $\{X \in B\}$.

La **loi d'une variable aléatoire discrète** est donnée par : $X(\Omega)$ et, pour tout $x \in X(\Omega)$, $p_x = \mathbb{P}(X = x)$. On a alors $\boxed{p_x \geq 0 \text{ et } \sum_x p_x = 1}$ et $\mathbb{P}(X \in B) = \sum_{x \in B} \mathbb{P}(X = x)$.

Le **Théorème de transfert** permet de calculer l'**espérance** de $\varphi(X)$ où $X(\Omega) \subset \mathbb{R}^d$, avec $\varphi : \mathbb{R}^d \to \mathbb{R}$: $\boxed{\mathbb{E}(\varphi(X)) = \sum_x \varphi(x)\mathbb{P}(X = x)}$ sous réserve de convergence absolue de cette série. En particulier, si X et Y sont des variables aléatoires discrètes réelles et $\varphi : \mathbb{R}^2 \to \mathbb{R}$, alors

$$\mathbb{E}(\varphi(X, Y)) = \sum_{x,y} \varphi(x, y)\mathbb{P}(X = x, Y = y).$$

Si X est une variable aléatoire discrète réelle, on a, pour les applications $\varphi : \mathbb{R} \to \mathbb{R}$ suivantes :

• $\varphi(x) = x$: $\boxed{\mathbb{E}(X) = \sum_x x\mathbb{P}(X = x) = m}$ (espérance mathématique de X ou moyenne de X) ;
• $\varphi(x) = (x - m)^2$ avec $m = \mathbb{E}(X)$, $\mathbb{E}((X - m)^2) = \sum_x (x - m)^2\mathbb{P}(X = x) = \sum_x x^2\mathbb{P}(X = x) - m^2 =$ $\boxed{\mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \mathbb{V}\text{ar}(X)}$ (variance de X) ;
• si $\varphi(x) = e^{itx}$ avec $t \in \mathbb{R}$ on obtient la **fonction caractéristique** $\varphi_X(t) = \mathbb{E}\left(e^{itX}\right) = \sum_x e^{itx}\mathbb{P}(X = x)$.
• si $X(\Omega) \subset \mathbb{N}$ et $\varphi(x) = s^x$ on obtient la fonction génératrice $\phi_X(s) = \mathbb{E}(s^X) = \sum s^k\mathbb{P}(X = k)$: c'est une série entière de rayon $R \geq 1$ et on a

$$\mathbb{E}(X) = G'_X(1-) \ \ ; \ \ G''_X(1-) = \mathbb{E}(X(X-1)) \ \ ; \ \ \mathbb{V}\text{ar}(X) = G''_X(1-) + G'_X(1-) - (G'_X(1-))^2.$$

Propriétés : $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$ et $\mathbb{V}\text{ar}(aX + b) = a^2\mathbb{V}\text{ar}(X)$. Écart-type de X : $\sigma_X = \sqrt{\mathbb{V}\text{ar}(X)}$.

Dans la suite, on prend $d = 2$ pour simplifier.

Lois marginales d'un couple discret (X, Y) : si $p_{x,y} = \mathbb{P}(X = x, Y = y)$ pour $x \in X(\Omega)$ et $y \in Y(\Omega)$, alors $p_{x.} = \mathbb{P}(X = x) = \sum_y p_{x,y}$ et $p_{.y} = \mathbb{P}(Y = y) = \sum_x p_{xy}$.
Indépendance : X et Y sont indépendantes si et seulement si $p_{x,y} = p_{x.}p_{.y}$ pour tout (x, y).

Covariance : $\boxed{\mathbb{C}\text{ov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}$.

Propriétés :
• $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$; $\mathbb{C}\text{ov}(X, X) = \mathbb{V}\text{ar}(X)$ et $\mathbb{C}\text{ov}(aX + b, cY + d) = ac\ \mathbb{C}\text{ov}(X, Y)$;
• $\mathbb{V}\text{ar}(aX + bY) = a^2\mathbb{V}\text{ar}(X) + b^2\mathbb{V}\text{ar}(Y) + 2ab\ \mathbb{C}\text{ov}(X, Y)$;
• Si X et Y sont indépendantes, alors $\mathbb{C}\text{ov}(X, Y) = 0$, mais la réciproque est fausse en général !

Probabilités conditionnelles : On s'intéresse à la probabilité que A ait lieu, sachant que B a lieu : ceci revient à remplacer Ω par B et A par $A \cap B$: si $\mathbb{P}(B) \neq 0$, $\boxed{\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}}$.

A et B indépendants (*i.e.* $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$) équivaut à $\mathbb{P}(B) = \mathbb{P}(B \mid A)$ ou à $\mathbb{P}(A) = \mathbb{P}(A \mid B)$: "la connaissance de B n'apporte aucune information pour déterminer la probabilité de A".

Probabilités totales : Lorsque $\Omega = \bigcup_n E_n$, avec les E_n disjoints deux à deux,

$$\mathbb{P}(A) = \sum_n \mathbb{P}(A \cap E_n) = \sum_n \mathbb{P}(A|E_n)\mathbb{P}(E_n)$$

Loi conditionnelle : Si $x \in X(\Omega)$ fixé, la loi $\mathbb{P}_Y(\,\cdot \mid X = x)$ est déterminée, pour tout $y \in Y(\Omega)$ par :

$$\mathbb{P}_Y(\{y\} \mid X = x) = \mathbb{P}(Y = y \mid X = x) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(X = x)} = \frac{p_{xy}}{p_{x.}}.$$

Espérance conditionnelle de Y à $X = x$: espérance d'une variable aléatoire de loi $\mathbb{P}_Y(\cdot \mid X = x)$. C'est un **réel** fonction de x. Plus précisément, si X et Y sont des variables aléatoires réelles discrètes, pour $x \in X(\Omega)$,

$$\mathbb{E}(Y \mid X = x) = \sum_{y \in Y(\Omega)} y\mathbb{P}(Y = y \mid X = x), \text{ sous réserve de convergence absolue.}$$

$\mathbb{E}(Y \mid X)$ est la **variable aléatoire** $\varphi(X)$ fonction de X où φ est la fonction réelle définie par $\varphi(x) = \mathbb{E}(Y \mid X = x)$.

Espérance totale : Si Y admet une espérance, $\mathbb{E}(Y \mid X)$ aussi et alors $\boxed{\mathbb{E}(\mathbb{E}(Y \mid X)) = \mathbb{E}(Y)}$.

Propriété : $\mathbb{E}(\psi(X)Y \mid X) = \psi(X)\mathbb{E}(Y \mid X)$ pour toute fonction bornée ψ.

Corollaire de l'inégalité de Markov : Si $\varepsilon > 0$ et $\alpha > 0$ alors

$$\mathbb{P}(|X| > \varepsilon) \leq \frac{1}{\varepsilon^\alpha}\mathbb{E}(|X|^\alpha)$$

Pour $\alpha = 2$ cela donne l'**inégalité de Bienaymé–Tchebychev** $\mathbb{P}(|X - \mathbb{E}(X)| > \varepsilon) \leq \frac{1}{\varepsilon^2}\mathbb{V}\text{ar}(X)$.

1.9 Exercices

Les exercices précédés d'une flèche ↪ sont des exercices d'application directs du cours.

↪ **Exercice 1.1** *(Manipulation de l'espérance conditionnelle)*

1. Prouvez la Proposition 1.7.2.

2. Prouvez le Théorème 1.7.5.

3. Soit X, X' i.i.d. : donnez une expression de $\mathbb{P}(X' \leq X \mid X)$ faisant intervenir la fonction de répartition de X.

4. Montrez que $\mathbb{E}(X \mid \xi_A) = \xi_A \mathbb{E}(X \mid A) + (1 - \xi_A)\mathbb{E}(X \mid A^c)$ et plus généralement que $\mathbb{E}(Y \mid X) = \sum_{x \in X(\Omega)} \mathbb{E}(Y \mid X = x)\mathbb{1}\{X = x\}$.

5. Soit N à valeurs dans $\mathbb{N}^*$ et $X_i \geq 0$: sans repasser par le conditionnement par des évènements, prouvez directement que $\mathbb{E}(\sum_{k=1}^N X_i \mid N) = \sum_{k=1}^N \mathbb{E}(X_i \mid N)$.
Indication : On écrira $\sum_{k=1}^N X_i = \sum_{k \geq 1} X_i \mathbb{1}\{k \leq N\}$ *et on utilisera les Propositions 1.3.8 et 1.7.2.*

6. Soit N à valeurs dans $\mathbb{N}^*$ et les $X_i \geq 0$ i.i.d. et indépendantes de N : montrez que $\mathbb{E}(\prod_{k=1}^N X_i \mid N) = \mathbb{E}(X_1)^N$.

↪ **Exercice 1.2** *(Calcul d'espérances conditionnelles)*

1. Soit X et Y des variables aléatoires géométriques indépendantes de même paramètre $p \in (0, 1)$. Calculez la loi conditionnelle de X sachant $X + Y$ et déduisez-en $\mathbb{E}(X \mid X + Y)$.

2. Montrez plus généralement que si $X_1, \ldots, X_n$ sont i.i.d., alors $\mathbb{E}(X_i \mid X_1 + \cdots + X_n) = (X_1 + \cdots + X_n)/n$.
Indication. Que vaut $\sum_{k=1}^n \mathbb{E}(X_k \mid X_1 + \cdots + X_n)$ *?*

3. Soit X et Y des variables aléatoires de Poisson indépendantes et de paramètre respectif $a, b > 0$. Montrez que $X+Y$ suit une loi de Poisson de paramètre $a+b$ par calcul direct et en utilisant les fonctions génératrices, puis calculez la loi conditionnelle de X sachant $X + Y$ et déduisez-en $\mathbb{E}(X \mid X + Y)$.

4. Soit X et Y des variables aléatoires indépendantes et uniformément réparties sur $\{1, \ldots, n\}$: calculez la loi conditionnelle de $\max(X, Y)$ sachant $\min(X, Y)$.

↪ **Exercice 1.3** *(Théorème de la variance totale)*

1. On définit la variance conditionnelle $\mathbb{V}\mathrm{ar}(Y \mid X)$ par

$$\mathbb{V}\mathrm{ar}(Y \mid X) = \mathbb{E}\left[(Y - \mathbb{E}(Y \mid X))^2 \mid X\right].$$

Montrez que

$$\mathbb{V}\mathrm{ar}(Y \mid X) = \mathbb{E}(Y^2 \mid X) - [\mathbb{E}(Y \mid X)]^2 \text{ et que } \mathbb{V}\mathrm{ar}(Y) = \mathbb{E}\big[\mathbb{V}\mathrm{ar}\,(Y \mid X)\big] + \mathbb{V}\mathrm{ar}\big[\mathbb{E}\,(Y \mid X)\big]. \tag{1.15}$$

2. (*Application numérique*) Une étude statistique faite pour un magasin a montré que le nombre de clients par semaines est une variable aléatoire de moyenne 400 et d'écart-type 20, et que la dépense de chaque client est une variable aléatoire de moyenne 100 euros et d'écart-type 100 euros. Les dépenses des différents clients sont mutuellement indépendantes, et sont indépendantes du nombre de clients. Calculez la moyenne et l'écart-type du chiffre d'affaires réalisé par le magasin pendant une semaine (le chiffre d'affaires est la recette totale).

Exercice 1.4

1. Soit X à valeurs entières : en admettant que l'on peut intervertir dérivation et sommation, montrez que

$$\mathbb{E}(X) = \phi'_X(1) \quad \text{et} \quad \mathbb{V}\mathrm{ar}(X) = \phi''_X(1) + \phi'_X(1)(1 - \phi'_X(1)).$$

Exercice 1.5

On suppose que le nombre X_t d'accidents dans une ville sur une durée de t jours est une variable aléatoire de loi de Poisson de paramètre at. D'autre part, le nombre de personnes impliquées dans un accident suit une loi géométrique de paramètre $1/2$.

1. Soit Z_t la variable aléatoire égale au nombre de personnes impliquées dans un accident sur une durée de t jours : en supposant que le nombre de personnes impliquées dans un accident est indépendant du nombre d'accidents, montrez que $\phi_{Z_t}(z) = \phi_{X_t} \circ \phi_G(z)$ avec G une variable aléatoire que l'on précisera.

2. Déduisez-en en utilisant l'exercice précédent l'espérance et la variance de Z_t.

3. Retrouvez ce résultat en utilisant les théorèmes de l'espérance totale et de la variance totale.

Problème 1.6

Un QCM se compose de 20 questions à 4 réponses possibles dont une seule est correcte et vaut 1 point. Le programme de l'examen comporte 100 questions dont on tire aléatoirement les 20 de l'examen. On considère un candidat ayant appris $p \geq 20$ questions : pour une question de l'examen, si le candidat l'a apprise alors il obtient un point et sinon il choisit une réponse au hasard. Le but de ce problème est de calculer la loi, l'espérance et la variance de la note N de ce candidat. Pour cela on écrira $N = X + Y$ avec X le nombre de questions apprises par le candidat et figurant à l'examen.

1. Calculez la loi conditionnelle de Y sachant X et déduisez-en $\mathbb{E}(Y \mid X)$ et $\mathbb{V}\mathrm{ar}(Y \mid X)$.

2. Déduisez-en $\mathbb{E}(N)$ et $\mathbb{V}\mathrm{ar}(N)$ en fonction de $\mathbb{E}(X)$ et $\mathbb{V}\mathrm{ar}(X)$ en utilisant les théorèmes de l'espérance totale et de la variance totale.

3. En considérant tous les questionnaires possibles, montrez que pour $k \in \{0, \ldots, 20\}$ avec $k \geq p - 80$ on a

$$\mathbb{P}(X = k) = \frac{\binom{p}{k}\binom{100-p}{20-k}}{\binom{100}{20}}.$$

Déduisez-en l'espérance et la variance de X.
Indication. On pourra utiliser l'identité

$$\sum_{j=0}^{k} \binom{m}{j}\binom{n}{k-j} = \binom{m+n}{k}, \ k \leq m, n$$

obtenue en développant $(x+y)^m(x+y)^n = (x+y)^{m+n}$. *Pour calculer la variance, on pourra chercher à calculer* $\mathbb{E}(X(X-1))$.

4. *(Application numérique)* Que vaut $\mathbb{E}(N)$ pour $p = 50$? $p = 70$?

Chapitre 2

Variables aléatoires absolument continues

Nous introduisons dans ce chapitre une nouvelle classe de variables aléatoires : les variables aléatoires absolument continues par rapport à la mesure de Lebesgue, que nous appellerons par simplicité variables aléatoires absolument continues. Le message principal est le suivant : si vous avez bien compris le chapitre précédent sur les probabilités discrètes, il sera facile de comprendre ce chapitre. En effet,

> **Pour passer du cas discret au cas continu,**
> **il suffit de remplacer les sommes $\sum$ par des intégrales $\int$**
> **et les probabilités $\mathbb{P}_X(\{x\})$ par les densités $f_X(x)\mathrm{d}x$.**

Tout le chapitre précédent a été écrit de telle sorte que le maximum de résultats n'ait pas besoin de changement, ainsi, ce chapitre sera beaucoup plus concis que le précédent.

2.1 Motivation et de la nécessité des tribus

Il y a au moins deux raisons d'aller au-delà du cadre des probabilités discrètes du chapitre précédent. La première est qu'il existe des expériences et variables aléatoires qui ne peuvent pas être décrites par un univers dénombrable. Un premier exemple est une série infinie de pile ou face : une description naturelle est alors $\Omega = \{0,1\}^{\mathbb{N}}$. De manière équivalente, on aimerait pouvoir parler d'un nombre tiré au hasard dans l'ensemble $[0,1]$.

Une deuxième motivation vient de la volonté naturelle de vouloir parler de convergence de variables aléatoires (cf. Chapitre 3). Considérons par exemple U_n pour $n \in \mathbb{N}^*$ une variable aléatoire discrète répartie uniformément sur l'ensemble $\{k/n : k = 0, \dots, n\}$. Pour chaque $n \in \mathbb{N}^*$, U_n est bien une variable aléatoire discrète mais intuitivement, lorsque $n \to \infty$ la suite $(U_n, n \in \mathbb{N}^*)$ devrait converger (en un sens à préciser) vers une variable aléatoire uniformément répartie sur $[0,1]$, et en particulier qui n'est plus discrète. De même, le théorème de la limite centrale justifie l'introduction des variables aléatoires gaussiennes qui sont des variables aléatoires absolument continues.

D'un point de vue conceptuel, il n'y a pas beaucoup de différences entre les variables aléatoires discrètes traitées au chapitre précédent et les variables aléatoires absolument continues

que nous allons maintenant introduire. Comme mentionné précédemment, il s'agit essentiellement de remplacer les sommes par les intégrales. En fait, cette distinction est artificielle puisque les deux cas peuvent être unifiés dans le cadre de la théorie de la mesure.

D'un point de vue technique par contre, il y a certaines différences importantes. En particulier, bien que dans le cas discret tout ensemble est mesurable – ce qui correspond au fait que l'on ait considéré dans tout le chapitre précédent $\mathcal{F} = \mathcal{P}(\Omega)$ – cela n'est plus vrai dans le cas non-dénombrable.

Prenons par exemple $\Omega = [0, 1]$, qui décrirait une expérience aléatoire où l'on tire au hasard un point de l'intervalle unité. On voudrait une mesure sur $[0, 1]$ qui satisfait la Propriété (1.1) de σ-additivité mais aussi la relation $\mathbb{P}([a, b]) = b-a$ pour tout $0 \leq a \leq b \in 1$: il s'avère que l'on peut prouver qu'il n'est pas possible de construire une telle mesure qui soit définie sur $\mathcal{P}([0, 1])$. Le lecteur intéressé pourra par exemple consulter l'exemple de Vitali.

Pour cette raison, dès lors que Ω n'est plus dénombrable on va devoir restreindre l'ensemble de définition des mesures de probabilité que l'on va considérer : cela nous amène donc en premier lieu à revisiter la notion d'espace de probabilités. Encore une fois, cette différence est la plus importante d'un point de vue technique, mais conceptuellement, tout a été vu au chapitre précédent.

2.2 Espace de probabilités : le cas général

Comme expliqué précédemment, pour développer une théorie satisfaisante dans le cas où Ω n'est pas dénombrable il est nécessaire de restreindre l'ensemble de définition des mesures de probabilité. Néanmoins, il faut quand même imposer certaines conditions, par exemple le fait que l'union de deux ensembles mesurables reste mesurable afin de pouvoir parler de la probabilité de l'évènement A ou B ; de même, il faut que le complément d'un ensemble mesurable reste mesurable. En fait, il ne faut pas beaucoup plus pour définir une théorie satisfaisante.

Définition 2.2.1. Un ensemble $\mathcal{F} \subset \mathcal{P}(\Omega)$ de parties de Ω est appelée **tribu** si les trois conditions suivantes sont satisfaites :

1. $\Omega \in \mathcal{F}$;
2. si $A \in \mathcal{F}$, alors $A^c \in \mathcal{F}$;
3. si $A_n \in \mathcal{F}$ pour chaque $n \in \mathbb{N}$, alors $\cup_{n\in\mathbb{N}} A_n \in \mathcal{F}$.

Le couple $(\Omega, \mathcal{F})$ est appelé **espace mesurable** et $A \in \mathcal{F}$ un **ensemble mesurable**, ou encore **évènement**.

Lorsque Ω est dénombrable, on considèrera toujours la tribu $\mathcal{F} = \mathcal{P}(\Omega)$ et l'on retombe dans le cadre du Chapitre 1. Dans le cas où Ω n'est pas dénombrable par contre, plusieurs choix sont possibles : comme on l'a discuté, $\mathcal{P}(\Omega)$ est souvent trop gros et $\{\emptyset, \Omega\}$ est bien évidemment trop petit. Dans le cas où Ω est un espace topologique, ce qui sera le cas dans tout le reste de ce chapitre, il y a un choix canonique.

Définition 2.2.2. Supposons que Ω est munie d'une topologie $\mathcal{T}$. La **tribu borélienne**, notée $\mathcal{B}(\Omega)$, est la plus petite tribu qui contienne $\mathcal{T}$.

De manière générale, on peut prouver que pour tout ensemble $\mathcal{F} \subset \mathcal{P}(\Omega)$ de parties de Ω, il existe une unique tribu qui est la plus petite tribu qui contient $\mathcal{F}$: pour $\mathcal{F} = \mathcal{T}$ cela justifie la définition de la tribu borélienne. En outre et pour revenir à l'exemple précédent

de la mesure de Lebesgue sur $\mathbb{R}$, on peut prouver que sur $\mathbb{R}$, la tribu borélienne $\mathcal{B}(\mathbb{R})$ est strictement plus petite que l'ensemble des parties $\mathcal{P}(\mathbb{R})$.

Les définitions d'une mesure de probabilité et d'un espace de probabilité, données à la Définition (1.1.1), restent inchangées, et une mesure de probabilité continue à satisfaire les propriétés de base de la Proposition 1.1.1. Néanmoins, comme nous le verrons par la suite dans le cas général une mesure de probabilité sur un ensemble non-dénombrable ne peut pas être décrite par la famille $(p_\omega, \omega \in \Omega)$.

2.3 Variables aléatoires et loi : le cas général

Lorsque Ω est dénombrable, toute application définie sur Ω est une variable aléatoire, cf. Définition 1.2.1. Dans le cas général, il faut prendre plus de précautions. On rappelle que f^{-1} est l'application image réciproque de f, qui coïncide avec son inverse lorsque f est bijective.

Définition 2.3.1. Soient $(\Omega, \mathcal{F})$ et $(\Omega', \mathcal{F}')$ deux espaces mesurables.
Une application $f : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$ est dite **mesurable** si $f^{-1}(A') \in \mathcal{F}$ pour tout $A' \in \mathcal{F}'$.
Une application mesurable est appelée, dans le contexte probabiliste, **variable aléatoire**.

Cette définition explique maintenant pourquoi les ensembles $\mathcal{F}$ étaient systématiquement spécifiés dans le chapitre précédent, bien que ces ensembles étaient toujours pris comme étant l'ensemble des parties de l'ensemble Ω correspondant : en général, le fait d'être mesurable dépend des tribus considérées et puisque l'on a le choix de la tribu, il est nécessaire de spécifier tout l'espace mesurable et pas juste l'univers.

Cette définition est à rapprocher de la définition de continuité : on rappelle qu'une application est continue si l'image réciproque d'un ouvert est un ouvert. Ici, il suffit de remplacer continu(e) par mesurable, i.e., une application est mesurable si l'image réciproque d'un ensemble mesurable est un ensemble mesurable. En particulier, lorsque l'on considère les tribus boréliennes, on a le résultat suivant.

Théorème 2.3.1. *Soient Ω et Ω' deux espaces topologiques. Si $f : \Omega \to \Omega'$ est continue, alors $f : (\Omega, \mathcal{B}(\Omega)) \to (\Omega', \mathcal{B}(\Omega'))$ est mesurable.*

Dans le cas où Ω était discret, il n'y avait aucune question à se poser : tout était mesurable. Maintenant il faut faire attention, au moins conceptuellement, et on a la généralisation de la Proposition 1.2.1 suivante.

Proposition 2.3.2. *Soient Ω, Ω' et Ω'' des espaces topologiques et $X : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$ une variable aléatoire.*
Alors pour toute application mesurable $\varphi : (\Omega', \mathcal{F}') \to (\Omega'', \mathcal{F}'')$, l'application $\varphi \circ X : (\Omega, \mathcal{F}) \to (\Omega'', \mathcal{F}'')$ est une variable aléatoire.
Ainsi, si X et Y sont deux variables aléatoires à valeurs dans $\Omega' = \mathbb{R}$ muni de sa tribu borélienne, alors toute combinaison linéaire de X et Y est une variable aléatoire, ainsi que X^2, $\min(X, Y)$, etc.

Bien qu'il existe des applications non-mesurables,

Toutes les fonctions que l'on rencontrera dans le cadre de ce cours sont mesurables.

Ainsi, s'il est indispensable d'introduire les notions de mesurabilité pour définir une théorie cohérente, dans le cadre de ce cours on pourra sans problème ignorer ces subtilités et continuer à considérer que tous les ensembles et fonctions que vous serez amenés à considérer sont mesurables (et ils le seront effectivement). Nous ne passerons donc pas de temps à donner des critères simples pour vérifier qu'une application est mesurable.

La Définition 1.2.2 de la loi d'une variable aléatoire et le Théorème 1.2.2 (ainsi que sa preuve) montrant que la loi d'une variable aléatoire est une mesure de probabilité restent vrais dans ce cadre général.

2.4 Le cas discret revisité

Dans le chapitre précédent, une mesure de probabilité discrète et une variable aléatoire discrète ont été définies comme étant des mesure de probabilité et des applications définies sur ensemble dénombrable. Néanmoins, ce n'est pas une définition très satisfaisante car on a vu que l'univers Ω n'était pas si important : par exemple, on pourrait tout aussi bien décrire l'expérience d'un lancer de dé à l'aide d'un univers continu, par exemple $[0, 1]$, l'expérience n'en reste pas moins discrète. Nous présentons donc maintenant la définition générale de ces objets.

Définition 2.4.1. Soient $(\Omega, \mathcal{F})$ et $(\Omega', \mathcal{F}')$ deux espaces mesurables.
Une mesure de probabilité $\mathbb{P}$ sur $(\Omega, \mathcal{F})$ est dite **discrète** si son support est dénombrable, i.e., s'il existe S dénombrable tel que $\mathbb{P}(S) = 1$.
Une variable aléatoire $X : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$ est dite **discrète** si sa loi est discrète, i.e., s'il existe S dénombrable tel que $\mathbb{P}(X \in S) = 1$.

Puisque le support de la loi de X est inclus dans $X(\Omega)$, on obtient qu'une variable aléatoire dont l'ensemble image est dénombrable est discrète. En particulier, si Ω est dénombrable toute variable aléatoire est discrète et cette définition est donc cohérente avec celle du chapitre précédent.

2.5 Variables aléatoires absolument continues

2.5.1 Définition

Dans le cadre de ce cours, on se restreindra à une classe particulière de variables aléatoires : les variables aléatoires absolument continues par rapport à la mesure de Lebesgue, que l'on appellera simplement variables aléatoires absolument continues. A noter qu'on définit directement une variable aléatoire absolument continue à valeurs dans $\mathbb{R}^n$ pour $n \geq 1$.

Définition 2.5.1. Soit $(\Omega, \mathcal{F})$ un ensemble mesurable.
Une variable aléatoire $X : (\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ est dite **absolument continue** s'il existe une fonction mesurable $f_X : (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)) \to (\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$ telle que la loi $\mathbb{P}_X$ de X s'écrive sous la forme

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A) = \int_A f_X(x)\mathrm{d}x = \int_{\mathbb{R}^n} f_X(x)\mathbb{1}\{x \in A\}\,\mathrm{d}x, \ A \in \mathcal{B}(\mathbb{R}^n).$$

Si la fonction f_X vérifie cette relation, alors elle est appelée **densité** de la (loi de la) variable aléatoire X.

Par définition, on a $\int_{\mathbb{R}^n} f_X(x)\mathrm{d}x = \mathbb{P}(X \in \mathbb{R}^d)$ et donc

$$\int_{\mathbb{R}^n} f_X(x)\mathrm{d}x = 1.$$

Il ne faut pas beaucoup plus pour définir une densité, notamment, si f est positive, mesurable et que son intégrale vaut 1, alors l'application $A \in \mathcal{F} \mapsto \int_A f$ définit la loi d'une variable aléatoire absolument continue de densité f.

L'interprétation géométrique est claire : l'aire (ou le volume, en plus grande dimension) sous la densité f_X d'une variable aléatoire X représente la probabilité que X appartienne à l'ensemble sur lequel on intègre.

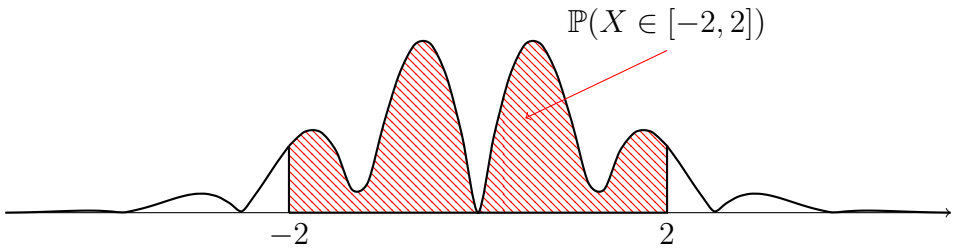


FIGURE 2.1 – L'aire sous la courbe entre -2 et 2 est égale à la probabilité $\mathbb{P}(X \in [-2, 2])$ que la variable aléatoire X soit entre -2 et 2.

Remarque 2.5.1. Lorsque l'on écrit $\int f_X(x)\mathrm{d}x$, il s'agit de l'intégration d'une fonction mesurable par rapport à la mesure de Lebesgue au sens de la théorie de la mesure : pour plus de détails, vous pouvez vous référer au Chapitre 8 du polycopié de mathématiques déterministes. Dans le cadre de ce cours, nous ne rencontrerons que des densités continues par morceaux (toute fonction continue par morceaux est mesurable, cf. Théorème 2.3.1) auquel cas l'intégrale de Lebesgue coïncide avec l'intégrale de Riemann.

Remarque 2.5.2. Dans la même veine, il n'y a pas unicité de la densité puisque si f et g coïncident partout sauf sur un ensemble de mesure nulle, alors pour tout A mesurable on a $\int_A f = \int_A g$. Néanmoins, avec un abus de langage nous continuerons à parler de "la" densité de X.

Une différence notable entre les cas discret et continu est que, alors que dans le cas discret toute fonction d'une variable aléatoire discrète est une variable aléatoire discrète (Proposition 1.2.1), cela n'est plus forcément le cas dans le cas absolument continu. D'une part, pour que $\varphi(X)$ reste une variable aléatoire il faut que φ soit mesurable. En outre, la fonction constante $\varphi(x) = x_0$ montre que $\varphi(X)$ n'est pas forcément absolument continue : on énonce maintenant un résultat qui découle du Théorème 8.3.16 du polycopié de mathématiques déterministes. On rappelle notamment que pour une application $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ bijective (sur un ensemble D), continûment différentiable ainsi que son inverse, $\mathrm{Jac}_x(\varphi)$ dénote sa matrice jacobienne prise au point x (cf. la Définition 3.2.1 et le Théorème 8.3.16 du polycopié de mathématiques déterministes) :

$$\mathrm{Jac}_x(\varphi) = \left(\frac{\partial \varphi_i}{\partial x_j}(x)\right)_{1 \le i \le m, 1 \le j \le n}.$$

Théorème 2.5.1. *On considère X à valeurs dans $\mathbb{R}^n$ et $\varphi : (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)) \to (\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$ mesurable. On suppose qu'il existe un ensemble ouvert $D \subset \mathbb{R}^n$ tel que :*

- *$f_X(x) = 0$ pour $x \notin D$;*
- *φ est bijective de D dans $\varphi(D)$;*
- *φ et φ^{-1} sont continûment différentiables sur D.*

Alors $\varphi(X)$ est absolument continue et sa densité est donnée par $f_{\varphi(X)}(y) = 0$ si $y \notin \varphi(D)$ et

$$f_{\varphi(X)}(y) = f_X\left(\varphi^{-1}(y)\right) \left|\det\left(\mathrm{Jac}_y(\varphi^{-1})\right)\right|, \ y \in \varphi(D).$$

Démonstration. Pour $A \in \mathcal{B}(\mathbb{R}^m)$ on a $\mathbb{P}\left(\varphi(X) \in A\right) = \mathbb{P}\left(X \in \varphi^{-1}(A)\right)$ par définition de φ^{-1}. Puisque φ est mesurable, $\varphi^{-1}(A) \in \mathcal{B}(\mathbb{R}^n)$ et donc la définition de la densité donne

$$\mathbb{P}\left(\varphi(X) \in A\right) = \int \mathbb{1}\left\{x \in \varphi^{-1}(A)\right\} f_X(x)\mathrm{d}x = \int \mathbb{1}\left\{\varphi(x) \in A\right\} f_X(x)\mathrm{d}x.$$

Puisque l'on peut se restreindre à $x \in D$ du fait que $f_X(x) = 0$ pour $x \notin D$, le changement de variable multi-dimensionnel $y = \varphi(x) \Leftrightarrow x = \varphi^{-1}(y)$ (Théorème 8.3.16 du polycopié de mathématiques déterministes) donne donc

$$\mathbb{P}\left(\varphi(X) \in A\right) = \int \mathbb{1}\left\{y \in A\right\} f_X\left(\varphi^{-1}(y)\right) \left|\det\left(\mathrm{Jac}_y(\varphi^{-1})\right)\right| \mathrm{d}y$$

ce qui prouve à la fois que $\varphi(X)$ est absolument continue et que sa densité est bien celle annoncée. ■

Dans le reste de ce chapitre, on utilisera les notations suivantes :

- Ω est un espace topologique et $\mathcal{F} = \mathcal{B}(\Omega)$;
- X, X', X'', Y, X_1, X_2, etc, sont des variables aléatoires absolument continues sur $(\Omega, \mathcal{F})$ et à valeurs dans $\mathbb{R}^n$, $\mathbb{R}^{n'}$, $\mathbb{R}^{n''}$, $\mathbb{R}^{n_Y}$, $\mathbb{R}^{n_1}$, $\mathbb{R}^{n_2}$, etc.

2.5.2 Densité : intuition et fausses idées

Le concept de densité est la seule vraie nouveauté du cas continu par rapport au cas discret : il est donc indispensable de bien comprendre ce que cette fonction signifie. Dans le cas discret on a

$$\mathbb{P}(X \in A) = \sum_{x \in A} \mathbb{P}_X(\{x\}) \qquad (X \text{ variable aléatoire discrète})$$

alors que dans le cas continu, on a par définition

$$\mathbb{P}(X \in A) = \int_{x \in A} f_X(x)\mathrm{d}x \qquad (X \text{ variable aléatoire absolument continue}).$$

On voit donc ici une application du premier encart de ce chapitre, que l'on répètera une nouvelle fois :

Pour passer du cas discret au cas continu,
il suffit de remplacer les sommes $\sum$ par des intégrales $\int$
et les probabilités $\mathbb{P}_X(\{x\})$ par les densités $f_X(x)\mathrm{d}x$.

Moralement, la valeur $f_X(x)$ prise au point x par la densité de X, variable aléatoire continue, joue le rôle analogue à la probabilité $\mathbb{P}_X(\{x\})$ de la probabilité de l'évènement élémentaire x sous la loi de X, variable aléatoire discrète. En particulier, on soulignera que

l'analogue de $\mathbb{P}_X(\{x\})$ pour X variable aléatoire discrète
N'EST PAS
$\mathbb{P}(X = x)$ pour X variable aléatoire absolument continue.

En effet, on a le résultat suivant qui découle directement de la définition de la densité.

Théorème 2.5.2. *Si X est une variable aléatoire absolument continue, alors pour tout $x \in \mathbb{R}^n$ on a $\mathbb{P}(X = x) = 0$.*

Démonstration. On vérifie facilement que $\{x\} \in \mathcal{B}(\mathbb{R}^n)$ et donc par définition de la densité on obtient $\mathbb{P}(X = x) = \int_{\{x\}} f(x')\mathrm{d}x'$ et puisque l'intégrale de Lebesgue de toute fonction mesurable sur un ensemble dénombrable est nulle on obtient bien $\mathbb{P}(X = x) = 0$. ■

Corollaire 2.5.3. *Une variable aléatoire absolument continue n'est pas discrète.*

Remarque 2.5.3. Il existe des variables aléatoires qui ne sont discrètes, ni absolument continues.

Remarque 2.5.4. Dans la même veine que le Théorème 2.5.2, et puisque la mesure de Lebesgue d'un espace vectoriel de dimension $r < n$ est nulle, on pourrait aussi montrer que si $X \in \mathbb{R}^n$ est absolument continue et que $E \subset \mathbb{R}^n$ est un espace vectoriel de dimension $\dim(E) < n$, alors $\mathbb{P}(X \in E) = 0$.

Remarque 2.5.5. Il suit de la remarque précédente que si X et Y sont absolument continues, alors (X, Y) n'est pas forcément absolument continue : il suffit pour voir cela de considérer $Y = X$. Néanmoins, si X et Y sont absolument continues et indépendantes, alors comme nous le verrons dans le Théorème 2.6.5, (X, Y) est bien absolument continue.

Le fait que $\mathbb{P}(X = x) = 0$ pour une variable absolument continue constitue une différence importante entre le cas discret et le cas continu : pour une variable aléatoire absolument continue à valeurs dans $\mathbb{R}$ par exemple, on a bien que $\mathbb{P}(X \in \mathbb{R}) = 1$ mais contrairement au cas discret, on n'a pas

$$\mathbb{P}(X \in \mathbb{R}) = \sum_{x \in X(\Omega)} \mathbb{P}(X = x) \qquad \boxed{\text{(FAUX ET HORRIBLE!!)}}$$

ni même

$$\mathbb{P}(X \in \mathbb{R}) = \int \mathbb{P}(X = x)\mathrm{d}x \qquad \boxed{\text{(FAUX ET (un peu moins) HORRIBLE!!)}}$$

En fait, sommer sur $X(\Omega)$ n'est même pas bien définie si $X(\Omega)$ n'est pas dénombrable et donc la première formule est encore plus horrible que la seconde. Il y a là un paradoxe

apparent : lorsque l'on fait une expérience aléatoire continue, on obtient bien un résultat. Néanmoins, la probabilité a priori d'obtenir ce résultat était nulle...

S'il ne faut pas directement interpréter la valeur de la densité comme une probabilité, le résultat suivant montre que l'on peut néanmoins donner un sens infinitésimal à cette interprétation. Ainsi, s'il est faux et horrible de dire que $f_X(x) = \mathbb{P}(X = x)$, il n'est pas inconcevable d'écrire $\mathbb{P}(X \approx x) \approx f_X(x)$ dans le sens suivant.

Théorème 2.5.4. *Si X est une variable aléatoire absolument continue à valeurs dans $\mathbb{R}$, alors pour tout point $x \in \mathbb{R}$ auquel f_X est continue,*

$$\mathbb{P}\big(X \in]x-\varepsilon, x+\varepsilon[\big) = 2f_X(x)\varepsilon + o(\varepsilon)$$

i.e.,

$$\lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} \mathbb{P}\big(X \in]x-\varepsilon, x+\varepsilon[\big) = f_X(x).$$

Démonstration. Il s'agit du théorème fondamental du calcul, cf. par exemple le Théorème 8.3.10 du polycopié de mathématiques déterministes. ■

2.5.3 Loi normale et autres lois usuelles

Une liste non-exhaustive de lois absolument continues inclut :

Loi uniforme : la loi uniforme sur un intervalle $[a, b] \subset \mathbb{R}$ avec $a \leq b$ correspond à choisir un point uniformément au hasard : sa densité vaut

$$f(x) = \frac{1}{b-a} \mathbb{1}\left\{x \in [a, b]\right\}, \ x \in \mathbb{R};$$

on peut plus généralement définir la loi uniforme sur un ensemble $A \subset \mathbb{R}^n$ de volume $\mathrm{Vol}(A) = \int_A \mathrm{d}x$ fini par $f(x) = \mathbb{1}\left\{x \in A\right\} / \mathrm{Vol}(A)$;

Loi normale sur $\mathbb{R}$: la loi normale (aussi appelée loi gaussienne) sur $\mathbb{R}$ de paramètre $(m, \sigma) \in \mathbb{R} \times]0, \infty[$ a pour densité

$$f(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right), \ x \in \mathbb{R}.$$

Il s'agit d'une loi de probabilité universelle qui apparaît dans le théorème central limite, cf. Théorème 3.4.7. Elle sera généralisée à la dimension supérieure dans le Chapitre 4 ;

Loi standard normale sur $\mathbb{R}$: il s'agit de la loi normale sur $\mathbb{R}$ de paramètre $m = 0$ et $\sigma = 1$, on parle aussi de loi normale centrée réduite ;

Loi exponentielle : la loi exponentielle de paramètre $\mu \in \mathbb{R}_+$ est la loi dont la densité vaut

$$f(x) = \mu e^{-\mu x} \mathbb{1}\left\{x \in \mathbb{R}_+\right\}, \ x \in \mathbb{R}.$$

Il s'agit de l'équivalent continu de la loi géométrique, cf. Proposition 3.4.6 ;

Loi Gamma : la loi Gamma de paramètre $\alpha, \beta > 0$ est la loi dont la densité est donnée par

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \mathbb{1}\left\{x \in \mathbb{R}_+\right\}, \ x \in \mathbb{R};$$

Loi Beta : la loi Gamma de paramètre $\alpha, \beta > 0$ est la loi dont la densité est donnée par

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}\mathbb{1}\left\{|x| \leq 1\right\}, \ x \in \mathbb{R}.$$

La loi normale joue un rôle important, il est notamment important de connaître sa fonction caractéristique. Le résultat suivant permet de ramener de nombreux calculs sur des lois normales au cas standard. A noter que les énoncés et démonstrations des deux résultats suivants anticipent légèrement et font appel aux notions de fonction de répartition et de fonction caractéristique dans le cas absolument continu, cf. Sections 2.6.1 et 2.6.9.

Proposition 2.5.5. *X suit une loi normale de paramètre (m, σ^2) si et seulement si $(X - m)/\sigma$ suit une loi normale standard.*

Démonstration. Soit X_{m,σ^2} qui suit une loi normale de paramètre (m, σ^2) et $Y_{m,\sigma^2} = (X_{m,\sigma^2} - m)/\sigma$: pour montrer le résultat, il suffit de prouver que Y_{m,σ^2} suit une loi normale centrée réduite, i.e., que $F_{Y_{m,\sigma^2}} = F_{Y_{0,1}}$. Par définition,

$$F_{Y_{m,\sigma^2}}(x) = \mathbb{P}\left(Y_{m,\sigma^2} \leq x\right) = \mathbb{P}\left(\frac{X_{m,\sigma^2} - m}{\sigma} \leq x\right) = \mathbb{P}\left(X_{m,\sigma^2} \leq \sigma x + m\right)$$

et en utilisant l'expression de la densité de X_{m,σ^2}, on obtient donc

$$F_{Y_{m,\sigma^2}}(x) = \int_{-\infty}^{\sigma x+m} \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-(y-m)^2/(2\sigma^2)} \mathrm{d}y.$$

Le changement de variable $u = (y - m)/\sigma$ donne donc

$$F_{Y_{m,\sigma^2}}(x) = \int_{-\infty}^{x} \frac{1}{(2\pi)^{1/2}} e^{-u^2/2} \mathrm{d}u = F_{Y_{0,1}}(x)$$

ce qui prouve le résultat. ■

Proposition 2.5.6. *La fonction caractéristique de la loi normale de paramètre (m, σ^2) est donnée par*

$$\varphi(x) = \exp\left(-\frac{x^2\sigma^2}{2} + ixm\right), \ x \in \mathbb{R}.$$

Démonstration. Soit X qui suit une loi normale de paramètre (m, σ^2) et $Y = (X - m)/\sigma$ qui suit une loi normale centrée réduite. Puisque

$$\varphi_X(x) = \mathbb{E}(e^{ixX}) = e^{ixm}\varphi_Y(\sigma x),$$

il suffit de montrer le résultat pour $m = 0$ et $\sigma = 1$, i.e., dans le cas standard. Dans ce cas, on a

$$\varphi_Y(x) = \frac{1}{(2\pi)^{1/2}} \int e^{ixy} e^{-y^2/2} \mathrm{d}y.$$

Le changement de variable $u = -y$ montre que $\varphi_Y(y) \in \mathbb{R}$, et donc

$$\varphi_Y(x) = \frac{1}{(2\pi)^{1/2}} \int \cos(xy) e^{-y^2/2} \mathrm{d}y.$$

Dérivant par rapport à x, on obtient

$$\varphi_Y'(x) = -\frac{1}{(2\pi)^{1/2}} \int y \sin(xy) e^{-y^2/2} \mathrm{d}y$$

et une intégration par parties donne $\varphi_Y'(x) = -x\varphi_Y(x)$. La résolution de cette équation différentielle avec la condition initiale $\varphi_Y(0) = 1$ donne la réponse. ■

2.6 Du discret au continu : les sommes deviennent des intégrales

Nous revisitons maintenant les notions introduites dans les Sections 1.3 à 1.5 dans le cas discret (fonction de répartition, espérance, variance, etc) : il s'agit essentiellement de remplacer $\sum$ par $\int$ et $\mathbb{P}_X(\{x\})$ par $f_X(x)\mathrm{d}x$.

2.6.1 Fonction de répartition

Si X est absolument continue à valeurs dans $\mathbb{R}^n$, la définition de sa fonction de répartition F_X est identique à celle du cas discret, cf. Définition 1.4.1 : $F_X(x) = \mathbb{P}(X \leq x)$. Lorsque X est à valeurs réelles et absolument continue, on énumère ci-dessous quelques propriétés de F_X : la seule différence avec le cas discret, cf. Proposition 1.3.1, est que F_X est continue et non plus constante par morceaux.

Proposition 2.6.1. *Si X absolument continue est à valeurs dans $\mathbb{R}^n$, alors sa fonction de répartition est F_X **continue** et croissante en chacune de ses variables. Si $n = 1$, F_X satisfait en outre*

$$\lim_{x\to-\infty} F_X(x) = 0 \quad \textit{et} \quad \lim_{x\to+\infty} F_X(x) = 1.$$

Démonstration. Cela découle directement de la formule $F_X(x) = \int_{-\infty}^{x} f_X$ et de résultats classiques d'analyse, cf. par exemple le Théorème 8.3.10 du polycopié de mathématiques déterministes. ■

Comme pour le cas discret (Corollaire 1.3.2), la fonction de répartition d'une variable aléatoire absolument continue caractérise sa loi : si X et Y sont absolument continues et $F_X = F_Y$, alors $\mathbb{P}_X = \mathbb{P}_Y$. On énonce maintenant un résultat qui, bien qu'apparemment anodin, a de nombreuses conséquences importantes.

Proposition 2.6.2. *Si $X \in \mathbb{R}$ et F_X est strictement croissante, alors la variable aléatoire $F_X(X)$ suit une loi uniforme sur $[0, 1]$.*

Démonstration. Pour prouver que $F_X(X)$ suit une loi uniforme, il suffit de prouver que sa fonction de répartition est celle de la loi uniforme, i.e., que

$$F_{F_X(X)}(x) = \begin{cases} 0 & \text{si } x \leq 0, \\ x & \text{si } 0 \leq x \leq 1, \\ 1 & \text{si } x \geq 1. \end{cases}$$

Puisque $F_{F_X(X)}(x) = \mathbb{P}\left(F_X(X) \leq x\right)$ par définition et que F_X est à valeurs dans $[0, 1]$, le résultat est évident pour $x \notin [0, 1]$. Pour $x \in [0, 1]$, on utilise le fait que F est une bijection croissante pour obtenir

$$\mathbb{P}\left(F_X(X) \leq x\right) = \mathbb{P}\left(X \leq F_X^{-1}(x)\right) = F_X\left(F_X^{-1}(x)\right) = x$$

ce qui prouve le résultat. ■

Remarque 2.6.1. Une conséquence fondamentale de ce résultat est qu'on peut générer n'importe quelle variable aléatoire à partir d'une variable uniforme : il suffit d'écrire $X = F_X^{-1}(U)$ avec $U = F_X(X)$. Bien qu'on n'a prouvé ce résultat que dans le cas où F_X est continue et strictement croissante, ce résultat reste en fait valable en tout généralité si l'on définit $F_X^{-1}(x) = \inf\{y \in \mathbb{R} : F_X(y) > x\}$.

2.6.2 Espérance : cas des variables absolument continues à valeurs dans $\mathbb{R}$

Pour définir l'espérance d'une variable absolument continue et à valeurs dans $\mathbb{R}_+$, on continue à remplacer $\sum$ par $\int$ et $\mathbb{P}_X(\{x\})$ par $f_X(x)\mathrm{d}x$, cf. la Définition 1.3.2 dans le cas discret.

Définition 2.6.1. Si X est à valeurs dans $\mathbb{R}_+$, alors son **espérance** $\mathbb{E}(X) \in \mathbb{R}_+ \cup \{\infty\}$ est définie de la manière suivante :

$$\mathbb{E}(X) = \int x f_X(x)\mathrm{d}x. \tag{2.1}$$

Le reste de la définition reste inchangé et, hormis le Théorème 1.3.6, tous les résultats de la Section 1.3.2 restent valables dans ce nouveau cadre. Quant au Théorème 1.3.6, il suffit encore une fois de remplacer $\sum$ par $\int$ et $\mathbb{P}_X(\{x\})$ par $f_X(x)\mathrm{d}x$ dans l'énoncé.

Théorème 2.6.3. *Supposons que X est à valeurs dans $\mathbb{R}^n$ et considérons $\varphi : (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mesurable. Si $\varphi(X) \geq 0$ ou $\mathbb{E}(|\varphi(X)|) < \infty$, alors*

$$\mathbb{E}\left(\varphi(X)\right) = \int \varphi(x) f_X(x)\mathrm{d}x. \tag{2.2}$$

Remarque 2.6.2. L'énoncé précédent cache une subtilité : ce n'est pas parce que X est absolument continue que $\varphi(X)$ l'est, et en particulier il n'est pas clair quel sens donner à $\mathbb{E}(\varphi(X))$. Dans le cas où $\varphi(X)$ est absolument continu, la preuve de ce résultat est la même que dans le cas discret en remplaçant les sommes par les intégrales. Dans le cas général, le résultat reste vrai et peut être vu, dans le cadre de ce cours, comme la définition de l'espérance d'une variable aléatoire qui peut s'écrire comme l'image d'une variable aléatoire absolument continue par rapport à une application mesurable.

2.6.3 Variance et écart-type : cas des variables absolument continues à valeurs dans $\mathbb{R}$

La Définition 1.3.3 et les Propositions 1.3.9 et 1.3.10 restent valables dans le cas absolument continu.

2.6.4 Inégalités de Markov, de Cauchy–Schwarz et de Bienaymé–Tchebychev

Les Théorèmes 1.3.11, 1.3.12 et 1.3.13 restent vrais dans le cadre absolument continu. La preuve du Théorème 1.3.13 utilise la version intégrable de l'inégalité de Cauchy–Schwarz, encore une fois il suffit de remplacer $\sum$ par $\int$ et $\mathbb{P}_X(\{x\})$ par $f_X(x)\mathrm{d}x$.

2.6.5 Lois marginales

Comme dans le cas discret, si $X = (X_1, \ldots, X_n)$ alors la loi de X_1 sera parfois qualifiée de **loi marginale**. Le Théorème 1.4.1 du cas discret se généralise au cas absolument continu en remplaçant, sans surprise, $\sum$ par $\int$ et $\mathbb{P}_X(\{x\})$ par $f_X(x)\mathrm{d}x$.

Théorème 2.6.4. *Soit $X = (X_1, \dots, X_n)$ avec $X_k \in \mathbb{R}^{n_k}$. On suppose que X est absolument continue de densité f_X. Alors X_1 est absolument continue et sa densité f_{X_1} est donnée par*

$$f_{X_1}(x_1) = \int_{x_2 \in \mathbb{R}^{n_2}, \dots, x_n \in \mathbb{R}^{n_n}} f_X(x_1, x_2, \dots, x_n)\, \mathrm{d}x_2 \cdots \mathrm{d}x_n, \ x_1 \in \mathbb{R}^{n_1}. \tag{2.3}$$

Démonstration. Soit $A \in \mathcal{B}(\mathbb{R}^{n_1})$: on a $\mathbb{P}(X_1 \in A) = \mathbb{P}(X_1 \in A, X_2 \in \mathbb{R}^{n_2}, \dots, X_n \in \mathbb{R}^{n_n})$ et par définition de la densité, on obtient donc

$$\mathbb{P}(X_1 \in A) = \int_{x_1 \in A, x_2 \in \mathbb{R}^{n_2}, \dots, x_n \in \mathbb{R}^{n_n}} f_X(x_1, x_2, \dots, x_n)\, \mathrm{d}x_2 \cdots \mathrm{d}x_n.$$

Le théorème de Fubini donne donc $\mathbb{P}(X_1 \in A) = \int_{x_1 \in A} f(x_1) \mathrm{d}x_1$ avec $f(x)$ donnée par le membre de droite de (2.3) pour $x_1 = x$. Par définition, cela identifie la densité de X_1 comme étant f ce qui prouve le résultat. ■

Ce résultat assure donc entre autre que si (X, Y) est absolument continue, alors chaque marginale l'est aussi. Néanmoins, il faut faire attention au fait que la réciproque n'est pas vraie : (X, X) n'est jamais absolument continue, même si X l'est. De manière plus générale, ce n'est pas parce que X et Y sont chacun absolument continus que le couple (X, Y) l'est.

Dans le cas d'une variable aléatoire absolument continue (X_1, X_2), le théorème précédent donne par exemple

$$\mathbb{P}_{X_1}(A) = \int_{x_1 \in A, x_2 \in \mathbb{R}^{n_2}} f_X(x_1, x_2) \mathrm{d}x_1 \mathrm{d}x_2, \ A \subset \mathbb{R}^{n_1}.$$

Cet exemple est en fait le plus général possible, puisqu'on peut s'y ramener en définissant $X_2 = (X_2, \dots, X_n)$ dans le théorème précédent. On retiendra donc que

La loi marginale de X_1 est obtenue en intégrant sur les valeurs possibles de X_2.

2.6.6 Indépendance

La Définition 1.4.2 de l'indépendance d'évènements et de variables aléatoires reste la même. Le Théorème 1.4.2 se généralise en remplaçant $\mathbb{P}_X(\{x\})$ par $f_X(x)$ de la manière suivante.

Théorème 2.6.5. *Soit X_1 et X_2 deux variables aléatoires absolument continues. Alors X_1 et X_2 sont indépendantes si et seulement si (X_1, X_2) est absolument continue et que la densité de la loi du couple (X_1, Y_2) est égale au produit des densités des lois marginales :*

$$f_{(X_1, X_2)}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2), \ x_1 \in \mathbb{R}^{n_1}, x_2 \in \mathbb{R}^{n_2}.$$

Les Propositions 1.4.3 et 1.4.4 restent inchangées.

2.6.7 Covariance : cas des variables absolument continues à valeurs dans $\mathbb{R}$

Les définitions et résultats de la Section 1.4.4 restent valables mot pour mot dans le cas de variables aléatoires absolument continues à valeurs dans $\mathbb{R}$.

2.6.8 Espérance, variance et covariance : généralisation au cas vectoriel et matriciel

Les définitions et notations de la Section 1.4.5 restent valables dans le cas absolument continu.

2.6.9 Fonction caractéristique, fonction génératrice et transformée de Laplace

Les définitions et résultats de la Section 1.5 restent valables dans le cas absolument continu, et les transformées des lois absolument classiques introduites en Section 2.5.3 sont données dans le Tableau B.2 en page 175. On remarquera en particulier que, en une dimension, la fonction caractéristique est donnée au vu du Théorème 2.6.3 par

$$\varphi_X(t) = \mathbb{E}\left(e^{itX}\right) = \int e^{itx} f_X(x)\mathrm{d}x$$

qui n'est rien d'autre, à une constante multiplicative près, que la transformée de Fourier de la fonction f_X vue en cours d'analyse fonctionnelle.

2.7 Conditionnement dans le cas absolument continu

Il reste maintenant à revisiter, dans le cas absolument continu, les notions de conditionnement introduites dans les Sections 1.6 et 1.7 dans le cas discret.

2.7.1 Conditionnement par rapport à un évènement

Les Définitions 1.6.1 et 1.6.2 et le Théorème 1.6.1 n'ont rien de spécifique au cas discret et restent vrais en toute généralité. Néanmoins, et on touche là à la limitation de l'approche de ce cours, il n'est pas évident de généraliser le Théorème 1.6.2 car avec l'approche utilisée, on ne sait pas quel sens donner à $\mathbb{E}(X\xi_A)$: sauf cas particulier, $X\xi_A$ n'est ni discrète ni absolument continue. Si on voit $\mathbb{E}(\,\cdot \mid A)$ comme l'opérateur d'espérance associée à la mesure de probabilité $\mathbb{P}(\,\cdot \mid A)$, on n'a que défini $\mathbb{E}(X \mid A)$ pour une variable aléatoire X absolument continue **sous** $\mathbb{P}(\,\cdot \mid A)$. Le résultat suivant exhibe un cas particulier très important où c'est bien le cas.

Lemme 2.7.1. *Soit $(X, Y) \in \mathbb{R}^{n_X} \times \mathbb{R}^{n_Y}$ absolument continue sous $\mathbb{P}$ de densité $f_{X,Y}$, et $B \in \mathcal{B}(\mathbb{R}^{n_Y})$ avec $\mathbb{P}(Y \in B) > 0$. Alors X est absolument continue sous $\mathbb{P}(\,\cdot \mid Y \in B)$ et sa densité conditionnelle est donnée par*

$$f_{X|Y\in B}(x) = \frac{1}{\mathbb{P}(Y \in B)} \int \mathbb{1}\{y \in B\} f_{X,Y}(x, y)\mathrm{d}y, \;\; x \in \mathbb{R}^{n_X}.$$

Démonstration. En utilisant successivement la définition de la probabilité conditionnelle, la définition de la densité de (X, Y) puis le théorème de Fubini, on obtient pour tout borélien

$A \in \mathcal{B}(\mathbb{R}^{n_X})$

$$\begin{aligned}\mathbb{P}(X \in A \mid Y \in B) &= \frac{\mathbb{P}(X \in A, Y \in B)}{\mathbb{P}(Y \in B)} \\ &= \frac{1}{\mathbb{P}(Y \in B)} \int \mathbb{1}\{x \in A, y \in B\} f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y \\ &= \int \mathbb{1}\{x \in A\} f_{X|Y\in B}(x)\mathrm{d}x\end{aligned}$$

ce qui prouve bien que $f_{X|Y\in B}$ est la densité de X sous $\mathbb{P}(\cdot \mid Y \in B)$. ■

Ce résultat permet donc de donner un sens à $\mathbb{E}(X \mid Y \in B)$. Par ailleurs, le Théorème 2.6.3 donne un sens à $\mathbb{E}(X; Y \in B) = \mathbb{E}(X\mathbb{1}\{Y \in B\})$ en considérant $\varphi(x,y) = x\mathbb{1}\{y \in B\}$. Un calcul élémentaire montre que, au terme multiplicatif $\mathbb{P}(Y \in B)$ près, ces deux espérances coïncident, généralisant ainsi le Théorème 1.6.2 au cas absolument continu pour un événement A de la forme $Y^{-1}(B) = \{Y \in B\}$.

Théorème 2.7.2. *Soit $(X,Y) \in \mathbb{R}^{n_X} \times \mathbb{R}^{n_Y}$ absolument continue sous $\mathbb{P}$ de densité $f_{X,Y}$, et $B \in \mathcal{B}(\mathbb{R}^{n_Y})$ avec $\mathbb{P}(Y \in B) > 0$. Si $X \geq 0$ ou X est intégrable, alors*

$$\mathbb{E}(X \mid Y \in B) = \frac{\mathbb{E}(X; Y \in B)}{\mathbb{P}(Y \in B)}.$$

2.7.2 Espérance et loi conditionnelles par rapport à une variable aléatoire

L'intuition derrière la notion d'espérance conditionnelle et décrite en Section 1.7.2 reste la même dans le cas absolument continu que dans le cas discret. Néanmoins, on ne peut plus directement définir $\mathbb{E}(Y \mid X) = h(X)$ avec $h(x) = \mathbb{E}(Y \mid X = x)$ car $\mathbb{P}(X = x) = 0$ (Théorème 2.5.2). Néanmoins, l'approche infinitésimale derrière la densité (Théorème 2.5.4) permet de comprendre comment définir l'espérance conditionnelle dans le cas absolument continu.

Intuitivement, on voudrait définir $h(x) \approx \mathbb{E}(Y \mid X \approx x)$ dans le sens suivant :

$$h(x) = \lim_{\varepsilon \to 0} \mathbb{E}(Y \mid X \in]x-\varepsilon, x+\varepsilon[).$$

En utilisant le Lemme 2.7.1 avec $B_\varepsilon =]x-\varepsilon, x+\varepsilon[$, on obtient

$$\mathbb{E}(Y \mid X \in]x-\varepsilon, x+\varepsilon[) = \int y f_{Y|X\in B_\varepsilon}(y)\mathrm{d}y$$

avec par définition

$$f_{Y|X\in B_\varepsilon}(y) = \frac{\int \mathbb{1}\{x-\varepsilon \leq x' \leq x+\varepsilon\} f_{X,Y}(x',y)\mathrm{d}x'}{\int \mathbb{1}\{x-\varepsilon \leq x' \leq x+\varepsilon\} f_X(x')\mathrm{d}x'}.$$

Lorsque $\varepsilon \downarrow 0$, le numérateur se comporte comme $2\varepsilon f_{X,Y}(x,y)$ et le dénominateur comme $2\varepsilon f_X(x)$, ce qui donne $f_{Y|X\in B}(y) \approx f_{X,Y}(x,y)/f_X(x)$ et finalement,

$$\mathbb{E}(Y \mid X \in]x-\varepsilon, x+\varepsilon[) \approx \int y f_{Y|X=x}(y)\mathrm{d}y \quad \text{avec} \quad f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

Pour tout x, $f_{Y|X=x}$ est **appelée densité de Y sachant $X = x$**. Par analogie avec le cas discret, on pourrait alors définir l'espérance conditionnelle $\mathbb{E}(Y \mid X) = h(X)$ avec $h(x) = 0$ si $f_X(x) = 0$ et

$$h(x) = \int y f_{Y|X=x}(y)\mathrm{d}y = \frac{1}{f_X(x)} \int y f_{X,Y}(x, y)\mathrm{d}y.$$

Néanmoins, nous allons adopter un point de vue légèrement plus direct. On commence par la remarque suivante.

Proposition 2.7.3. *On a* $\mathbb{P}(f_X(X) = 0) = 0$.

Ainsi, on peut supposer que $f_X(X) > 0$ puisque cela arrive avec probabilité un. Cela nous permet d'adopter la définition suivante.

Définition 2.7.1. Pour tout y, la densité aléatoire

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

est appelée densité de Y sachant X : c'est la densité de la loi conditionnelle de Y sachant X.

$f_{Y|X}$ induit donc une mesure de probabilité notée $\mathbb{P}(Y \in \cdot \mid X)$ et appelée **loi conditionnelle de Y sachant X** via la formule suivante :

$$\mathbb{P}(Y \in A \mid X) = \int \mathbb{1}\{y \in A\} f_{Y|X}(y)\mathrm{d}y.$$

On a alors le résultat suivant, qui généralise le théorème de transfert discret 1.7.5 au cas absolument continu et qui définit notamment directement $\mathbb{E}(Y \mid X)$.

Théorème 2.7.4. *Soit (X, Y) absolument continu et $\varphi : \mathbb{R}^{n_Y} \to \mathbb{R}$ mesurable : si $\varphi(Y) \geq 0$ ou $\mathbb{E}(|\varphi(Y)|) < \infty$, alors*

$$\mathbb{E}(\varphi(Y) \mid X) = \int \varphi(y) f_{Y|X}(y)\mathrm{d}y. \tag{2.4}$$

Avec ces définitions, les résultats de la Section 1.7.3 restent valables. Dans la Proposition 1.7.1, il faut juste remplacer $\sum$ par $\int$ et $\mathbb{P}_Z(\{z\})$ par $f_Z(z)\mathrm{d}z$, i.e., si $Y = \varphi(X, Z)$ avec (X, Y, Z) absolument continu et X et Z indépendantes, alors $\mathbb{E}(Y \mid X) = \int \varphi(X, z) f_Z(z)\mathrm{d}z$.

2.8 Limitations

Comme on l'a vu, la théorie des variables absolument continues présente certaines limitations intrinsèques qui rendent par exemple difficile de définir proprement l'espérance conditionnelle. Une autre limitation est que l'on sort très naturellement du cadre absolument continu. Le premier exemple, mentionné précédemment, consiste à considérer le couple (X, X) avec X absolument continue. Un autre exemple naturel vient du traitement du signal : imaginons qu'avec probabilité p, on reçoive un signal X modélisé par une loi normale, et qu'avec la probabilité complémentaire $1 - p$ on ne reçoive aucun signal. La variable aléatoire correspondante est alors εX avec ε une variable de Bernoulli, qui n'est ni discrète ni absolument continue.

Seule la théorie de la mesure permet de construire une théorie des probabilités complètement satisfaisante. Une introduction à cette théorie est donnée dans le cours électif “Approfondissement en mathématiques” proposée une année sur deux dans la séquence commune 1A/2A.

2.9 Fiche de synthèse

Remarque importante : Presque tout ce qui a été fait dans le cas discret (définitions, propriétés,...) est encore valable dans le cas continu, en remplaçant Σ par $\int$ et $\mathbb{P}(X = x)$ par $f_X(x)\mathrm{d}x$...

Si Ω n'est pas dénombrable, toutes ses parties ne sont pas intéressantes, d'où l'introduction de la notion de tribu.

Tribu $\mathcal{F}$ sur Ω : tout sous-ensemble de parties de Ω tel que :
i) $\Omega \in \mathcal{F}$;
ii) si $A \in \mathcal{F}$, alors $A^c \in \mathcal{F}$;
iii) si pour tout $n \in \mathbb{N}$, $A_n \in \mathcal{F}$, alors $\bigcup_{n=0}^{+\infty} A_n \in \mathcal{F}$.
Probabilité $\mathbb{P}$ sur $(\Omega, \mathcal{F})$: toute application $\mathbb{P}$ de $\mathcal{F}$ vers $[0, 1]$ telle que :
i) $\mathbb{P}(\Omega) = 1$;
ii) pour toute suite d'événements $A_n \in \mathcal{F}$, incompatibles deux à deux, on a : $\mathbb{P}\left(\bigcup_{n=0}^{+\infty} A_n\right) = \sum_{n=0}^{+\infty} \mathbb{P}(A_n)$.

Variable aléatoire X : application mesurable $X : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$ ($X^{-1}(B) \in \mathcal{F}$ pour tout $B \in \mathcal{F}'$)
Loi de X : $\mathbb{P}_X$ probabilité sur $(\Omega', \mathcal{F}')$ définie par $\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B))$, où $\mathbb{P}$ probabilité sur $(\Omega, \mathcal{F})$.
Le plus souvent, $(\Omega', \mathcal{F}') = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ où $\mathcal{B}(\mathbb{R}^d)$ est la tribu des boréliens.

Une variable $X : (\Omega, \mathcal{F}) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ est dite **absolument continue** s'il existe une fonction mesurable $f_X : (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \to (\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$, appelée **densité** telle que, pour tout $B \in \mathcal{B}(\mathbb{R}^d)$,

$$\boxed{\mathbb{P}_X(B) = \mathbb{P}(X \in B) = \int_B f_X(x)\mathrm{d}x = \int f_X(x)\mathbb{1}\{x \in B\}\,\mathrm{d}x}.$$

En particulier, toute densité satisfait $\int f_X(x)\mathrm{d}x = 1$.

Loi de $U = \varphi(X)$ où X est de densité f_X, nulle hors de D et h C^1 difféomorphisme de $D \subset \mathbb{R}^d$ sur $D' = \varphi(D)$:

$$\boxed{f_{\varphi(X)}(y) = f_X\left(\varphi^{-1}(y)\right)\left|\det\left(\mathrm{Jac}_y(\varphi^{-1})\right)\right|,\ y \in \varphi(D)}$$

En général, on veut la loi de $\varphi(X, Y)$ où $\varphi : \mathbb{R}^2 \to \mathbb{R}$ (par exemple la loi de $X + Y$) : on peut commencer à chercher la loi de $(U, V) = (X, \varphi(X, Y))$ puis on intégre pour obtenir la deuxième marginale.

Le **Théorème de transfert** permet de calculer l'**espérance** de $\varphi(X)$ où $X(\Omega) \subset \mathbb{R}^d$, avec $\varphi : \mathbb{R}^d \to \mathbb{R}$: $\boxed{\mathbb{E}(\varphi(X)) = \int \varphi(x)\, f_X(x)\mathrm{d}x}$ sous réserve d'intégrabilité de $x \mapsto \varphi(x)f_X(x)$.
Fonction caractéristique : si $t \in \mathbb{R}^d$, pour $\varphi(x) = e^{i\langle t,x\rangle}$, $\varphi_X(t) = \mathbb{E}\left(e^{i\langle t,X\rangle}\right) = \int e^{i\langle t,x\rangle}\, f_X(x)\mathrm{d}x$.
Fonction de répartition : $F_X : x \in \mathbb{R}^d \mapsto \mathbb{P}(X_1 \leq x_1, \cdots, X_d \leq x_d)$.

Dans la suite, on prend $d = 2$ pour simplifier.

Lois marginales : $F_X(x) = \lim_{y \to +\infty} F_{X,Y}(x, y)$; $F_Y(y) = \lim_{x \to +\infty} F_{X,Y}(x, y)$.
Si (X, Y) absolument continu a pour densité $f_{X,Y}$, alors X et Y sont absolument continues, de densités respectives f_X et f_Y définies par $f_X(x) = \int f_{X,Y}(x, y)\mathrm{d}y$ et $f_Y(y) = \int f_{X,Y}(x, y)\mathrm{d}x$.

Cas particulier important où $d = 1$: $F_X(x) = \int_{-\infty}^{x} f_X(s)\mathrm{d}s$, F_X est continue, croissante sur $\mathbb{R}$, de classe $\mathcal{C}^1$ presque partout, avec alors $F_X' = f_X$. On a aussi $\mathbb{P}(X = x) = 0$ pour tout $x \in \mathbb{R}$ et $\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\mathrm{d}x$.

Loi de $U = \varphi(X)$: utiliser les fonctions de répartition. $F_U(u) = \mathbb{P}(\varphi(X) \leq u)$, à exprimer à l'aide de F_X puis dériver ...

Loi conditionnelle dans le cas continu : Pour $x \in X(\Omega)$ fixé tel que $f_X(x) \neq 0$, $\mathbb{P}_Y(\cdot \mid X = x)$ est une loi absolument continue de densité $f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}$.

Remarque : Le réel x étant fixé, il ne doit pas y avoir d'indicatrice dans $f_X(x)$ et dans $f_{X,Y}(x,y)$, c'est y qui doit être exprimé en fonction de x et non l'inverse.

Espérance conditionnelle de Y à $X = x$: espérance d'une variable de loi $\mathbb{P}_Y(\cdot \mid X = x)$. C'est un **réel** fonction de x. Plus précisément, si (X, Y) est absolument continu,

$$\mathbb{E}(Y \mid X = x) = \int y\, f_{Y|X=x}(y)\mathrm{d}y.$$

$\mathbb{E}(Y \mid X)$ est la **variable aléatoire** $\varphi(X)$ fonction de X où φ est la fonction réelle définie par $\varphi(x) = \mathbb{E}(Y \mid X = x)$.

2.10 Exercices

Les exercices précédés d'une flèche $\hookrightarrow$ sont des exercices d'application directs du cours.

$\hookrightarrow$ **Exercice 2.1**

1. Montrez en utilisant le théorème de transfert que $\mathbb{P}(f_X(X) = 0) = 0$.

$\hookrightarrow$ **Exercice 2.2** *(Calcul de densités)*

1. Montrez que
$$f_{aX+b}(x) = \frac{1}{|a|} f_X\left(\frac{x-b}{a}\right).$$

2. Montrez que $f_{|X|}(x) = (f_X(x) + f_X(-x))\mathbb{1}\{x \geq 0\}$.

3. Montrez que si X et Y sont indépendantes, alors $f_{X+Y}(z) = \int f_X(x) f_Y(z-x)\mathrm{d}x$.

4. Montrez que si X et Y sont indépendantes, alors
$$f_{\max(X,Y)}(z) = f_X(z)\mathbb{P}(Y \leq z) + f_Y(z)\mathbb{P}(X \leq z)$$

5. Soit U uniformément répartie sur $[0,1]$ et $c > 0$: calculez la loi de $-c\ln(U)$.

6. Soit U_1, U_2 indépendantes et uniformément réparties sur $[0,1]$: calculez la loi du couple
$$\left(\cos(2\pi U_1)\sqrt{-2\ln(U_2)}, \sin(2\pi U_1)\sqrt{-2\ln(U_2)}\right)$$
et déduisez en que ces deux variables sont i.i.d. standard normales.

$\hookrightarrow$ **Exercice 2.3** *(Loi normale)*
On rappelle que la densité de la loi normale de paramètre $(m, \sigma^2) \in \mathbb{R} \times (0, \infty)$ est donnée par
$$x \in \mathbb{R} \mapsto \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right).$$

1. Par un calcul direct, calculez la moyenne et la variance de X.

On suppose dorénavant que $m = 0$ et $\sigma = 1$.

2. Montrez que $\mathbb{E}(X^k) = 0$ pour $k \in \mathbb{N}$ impair.

3. Montrez par récurrence que $\mathbb{E}(X^{2k}) = (2k-1) \times (2k-3) \times \cdots \times 3 \times 1$ pour $k \in \mathbb{N}$.

4. Montrez que pour $a < b$,
$$\mathbb{E}(X \mid a < X < b) = \frac{f_X(a) - f_X(b)}{F_X(b) - F_X(a)}.$$

5. Soient X et Y deux variables aléatoires normales indépendantes : quelle est la loi de $X + Y$?
Indication : vous pourrez considérer les fonctions caractéristiques.

$\hookrightarrow$ **Exercice 2.4** *(Lois Gamma et Beta)*
La loi Gamma de paramètre (α, λ) est la loi de probabilité sur $\mathbb{R}$ dont la densité est donnée par
$$f^{\Gamma}_{\alpha,\lambda}(x) = c^{\Gamma}_{\alpha,\lambda} e^{-\lambda x} x^{\alpha-1} \mathbb{1}\{x \geq 0\}.$$

La loi Beta de paramètre (α, β) est la loi de probabilité sur $\mathbb{R}$ dont la densité est donnée par
$$f^{\beta}_{\alpha,\beta}(x) = c^{\beta}_{\alpha,\beta} x^{\alpha-1}(1-x)^{\beta-1} \mathbb{1}\{0 \leq x \leq 1\}.$$

1. Pour quelles valeurs des paramètres ces lois sont-elles bien définies ? Montrez alors que

$$c^{\Gamma}_{\alpha,\lambda} = \frac{\lambda^\alpha}{\Gamma(\alpha)} \quad \text{avec} \quad \Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}\mathrm{d}t$$

et que

$$c^{\beta}_{\alpha,\beta} = \frac{1}{B(\alpha,\beta)} \quad \text{avec} \quad B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x.$$

2. Montrez que $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ et que $\Gamma(\alpha)\Gamma(\beta) = B(\alpha,\beta)\Gamma(\alpha+\beta)$.
Indication. Pour la deuxième identité on pourra partir du membre de gauche et faire intervenir le difféomorphisme $\varphi^{-1} : (z,t) \in \mathbb{R}_+ \times [0,1] \mapsto (zt, z(1-t)) \in \mathbb{R}_+^2$ pour changer de variable.

3. Calculez l'espérance, la variance et la transformée de Laplace de chacune de ces lois.

4. Soit U et V deux variables aléatoires indépendantes suivant des lois $\Gamma(\alpha,1)$ et $\Gamma(\beta,1)$, respectivement : montrez que $U/(U+V)$ suit une loi $\mathrm{Beta}(\alpha,\beta)$.
Indication. On pourra utiliser le même changement de variable qu'à la question 2.

↪ **Exercice 2.5** *(Loi exponentielle)*
Soit $E_1, E_2, \ldots$ des variables exponentielles i.i.d..

1. Calculez la loi de $E_1 + \cdots + E_n$.
Indication. On pourra calculer la transformée de Laplace puis l'identifier à l'aide de l'exercice précédent.

2. A l'aide de la question précédente et du théorème de l'espérance totale, calculer la loi de $E_1 + \cdots + E_G$ où G est une variable géométrique indépendante.
Indication. On pourra calculer puis identifier la transformée de Laplace.

3. Calculez la loi de $E_1/(E_1+E_2)$.

↪ **Exercice 2.6**
Soit (X,Y) un couple aléatoire de densité f définie par :

$$f(x,y) = \begin{cases} \sqrt{\dfrac{x}{y}} & \text{si } (x,y) \in \Delta = \{(x,y) \ : \ 0 < y \leq x \leq 1\}, \\ 0 & \text{sinon.} \end{cases}$$

1. Représentez Δ, puis vérifiez que f est bien une densité et déterminez les lois marginales de X et de Y.

2. Les variables aléatoires X et Y sont-elles indépendantes ? Calculez $\mathbb{E}(X)$, $\mathbb{E}(Y)$, $\mathbb{E}(XY)$ et déduisez-en puis $\mathbb{C}\mathrm{ov}(X,Y)$.

3. Déterminez la loi du couple $(X, Y/X)$ et en déduire la loi de $U = Y/X$. Les variables aléatoires X et U sont-elles indépendantes ?

4. Trouvez la loi conditionnelle de Y sachant X. Déduisez-en $\mathbb{E}(Y \mid X)$ et retrouvez $\mathbb{E}(Y)$.

5. Calculez $\mathbb{P}(Y < X/2)$ puis déterminez la loi conditionnelle de X sachant Y.

6. Déduisez-en que $\mathbb{E}(X \mid Y) = \frac{3}{5}\frac{1-Y^{5/2}}{1-Y^{3/2}}$ et retrouvez $\mathbb{E}(X)$.

↪ **Exercice 2.7**
Soit X une variable aléatoire de densité f définie par

$$f_X(x) = \frac{1}{\sqrt{\pi x}}\, e^{-x} \mathbb{1}\{x > 0\}$$

et Y une variable aléatoire dont la loi conditionnelle sachant $X = x$ est la loi normale de moyenne 0 et de variance $\frac{1}{2x}$.

1. Rappelez la définition de $f_{Y|X=x}$ puis calculez la loi du couple (X,Y) et celle de Y.

2. Déterminez la loi conditionnelle de X sachant $Y = y$ et déduisez-en $\mathbb{E}(X \mid Y)$. Calculez $\mathbb{E}(X)$ et déduisez de ce qui précéde, sans calcul, la valeur de $\int_0^{+\infty} \frac{\mathrm{d}t}{(1+t^2)^2}$.

↪ **Exercice 2.8**

Deux personnes conviennent de se retrouver entre 17h et 18h, et on suppose que chaque personne arrive à un instant choisi uniformément au hasard dans cet intervalle.

1. Si la première personne arrivée n'attend pas plus de 10 minutes, quelle est la probabilité qu'elles se rencontrent ?

2. Soit T_i, $i = 1, 2$ l'heure d'arrivée de la personne i. Calculez la loi du temps d'attente $|T_2 - T_1|$ et retrouvez le résultat de la question précédente.

Problème 2.9

Soit $E_1, E_2, \ldots$ des variables i.i.d. distribuées selon la loi exponentielle de paramètre μ : le but de ce problème est de prouver que $\max(E_1, \ldots, E_n)$ et $E_1 + E_2/2 + \cdots + E_n/n$ sont égales en distribution. La propriété fondamentale qui permet de prouver cette égalité est la propriété suivante, dite d'absence de mémoire.

1. Soit $x \geq 0$: montrez que $E_1 - x$ sachant $E_1 \geq x$ est égale en loi à E_1.

2. Montrez que pour tout $\alpha > 0$, E_1/α suit la loi exponentielle de paramètre $\alpha\mu$.

3. Montrez que pour tout $\lambda, \lambda' > 0$, on a

$$\mathbb{E}\left(e^{-\lambda E_2 - \lambda'(E_1 - E_2)} \mid E_2 \leq E_1\right) = \mathbb{E}(e^{-\lambda E_1/2})\mathbb{E}(e^{-\lambda' E_1}).$$

4. En déduire que la loi de $(E_2, E_1 - E_2)$ conditionnellement à $E_2 \leq E_1$ est la même que celle de $(E_2/2, E_1)$, puis que $\max(E_1, E_2)$ et $E_1 + E_2/2$ ont même loi.

5. Généralisez les calculs précédents pour montrer que la loi de $(E_n, E_1 - E_n, \ldots, E_{n-1} - E_n)$ conditionnellement à $E_n = \min_{k=1,\ldots,n} E_k$ est la même que celle de $(E_n/n, E_1, \ldots, E_{n-1})$, puis concluez par récurrence que $\max(E_1, \ldots, E_n)$ et $E_1 + E_2/2 + \cdots + E_n/n$ ont même loi.

6. En déduire que

$$\mathbb{E}(\max(E_1, \ldots, E_n)) = \frac{1}{\mu}\sum_{k=1}^{n}\frac{1}{k} \quad \text{et} \quad \mathbb{V}\mathrm{ar}(\max(E_1, \ldots, E_n)) = \frac{1}{\mu^2}\sum_{k=1}^{n}\frac{1}{k^2}.$$

7. Quelle est la loi de $\min_{k=1,\ldots,n} E_k$?

Chapitre 3

Théorèmes limites

Dans ce chapitre nous introduisons les deux résultats les plus importants de la théorie des probabilités, avec des ramifications dans de nombreux domaines scientifiques et en ingénierie : la loi des grands nombres et le théorème central limite. Ces deux résultats peuvent être compris à l'aide d'une série infinie de pile ou face, et nous commencerons par présenter quelques résultats de simulation qui illustrent les concepts clef de ce chapitre.

3.1 Série de pile ou face

Considérons une série infinie de pile ou face : on définit

$$X_k = \begin{cases} 1 & \text{si on obtient pile au } k\text{-ième lancer,} \\ 0 & \text{sinon.} \end{cases}$$

On suppose que la pièce est non biaisée, la suite $(X_k, k \in \mathbb{Z}_+)$ est donc i.i.d. et la loi commune aux X_k est la loi de Bernoulli de paramètre $1/2$. On s'intéresse en particulier au comportement asymptotique quand $n \to \infty$ de la moyenne empirique $\bar{X}_n$ définie par

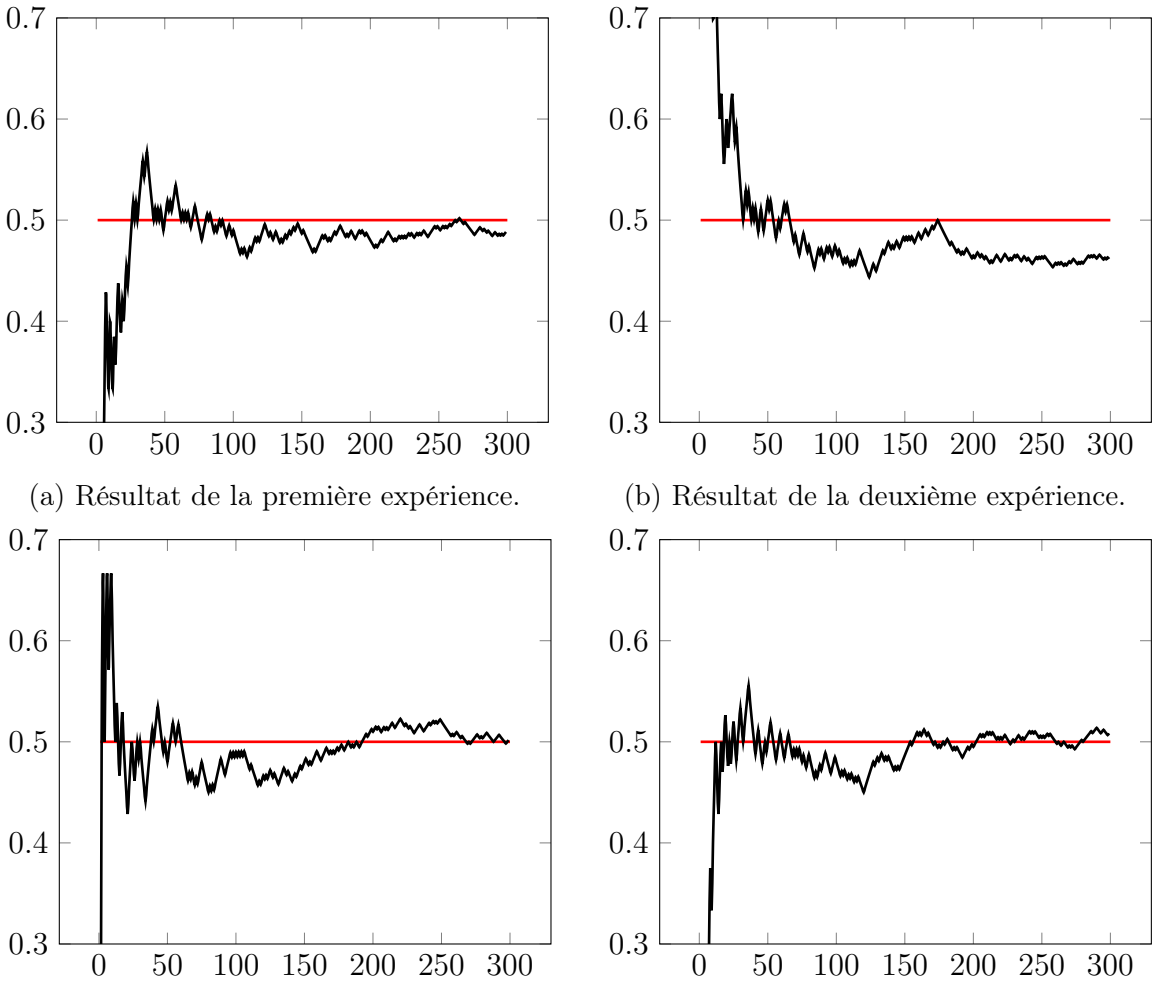
$$\bar{X}_n = \frac{1}{n} \sum_{k=1}^{n} X_k, \ \ n \in \mathbb{N}^*.$$

Dans les trois pages suivantes, nous présentons des résultats de simulation qui illustrent trois résultats fondamentaux :

Loi forte des grands nombres (Figure 3.1) : Pour quatre réalisations de l'expérience, on trace l'évolution de $\bar{X}_n$ pour $n = 1, \ldots, 300$. Ces résultats suggèrent que, **pour chaque expérience**, on a $\bar{X}_n \to 1/2$ lorsque $n \to \infty$, ce qui est conforme à l'intuition ;

Loi du logarithme itéré (Figure 3.2) : Pour l'une des quatre réalisations de l'expérience (de la Figure 3.1b) on n'a pas $\bar{X}_{300} \approx 1/2$. Deux explications sont possibles : $\bar{X}_n$ ne converge pas vers $1/2$, ou bien il y a bien convergence mais la suite n'a pas encore convergé au bout de 300 itérations. Pour décider, on s'intéresse donc au deuxième ordre de $\bar{X}_n$, i.e., on voudrait un développement asymptotique du genre $\bar{X}_n = 1/2 + \varepsilon_n + o(\varepsilon_n)$: la Figure 3.2 montre qu'une telle suite ε_n n'existe pas ! !

Théorème central limite (Figure 3.3) : Au lieu d'exiger un développement asymptotique du type $\bar{X}_n = 1/2 + \varepsilon_n + o(\varepsilon_n)$, on va seulement exiger un contrôle en probabilité : on va chercher ε_n tel que, pour tout $a \leq b$ réels, la probabilité $\mathbb{P}(a \leq (\bar{X}_n - 1/2)/\varepsilon_n \leq b)$ converge lorsque $n \to \infty$ vers une quantité non-triviale.



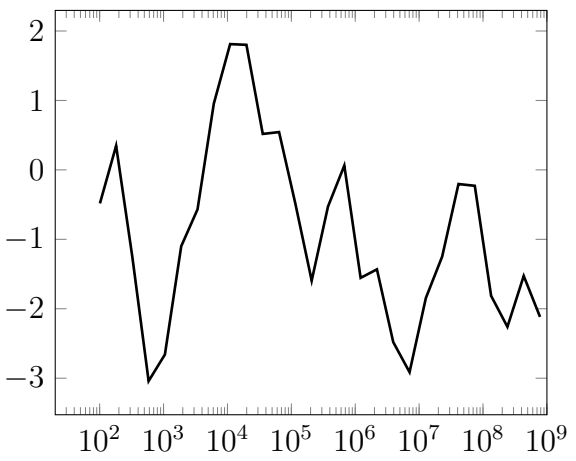
(a) Résultat de la première expérience.

(b) Résultat de la deuxième expérience.

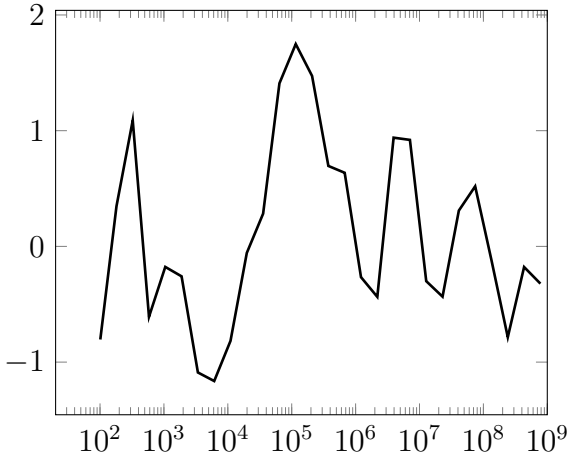
(c) Résultat de la troisième expérience.

(d) Résultat de la quatrième expérience.

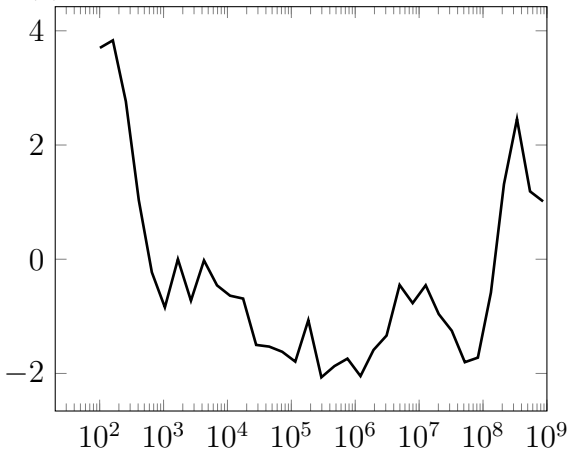
FIGURE 3.1 – Chaque courbe représente l'évolution de la moyenne empirique $(\bar{X}_n)$ pour une série de pile ou face : la ligne rouge correspond à la valeur asymptotique attendue 1/2. Pour les expériences a, c et d la moyenne empirique semble converger vers 1/2, ce qui est le résultat attendu. Pour l'expérience b la convergence semble plausible mais n'est pas évidente : en fait, elle a bien lieu mais il faudrait considérer un horizon temporel plus grand pour s'en convaincre.



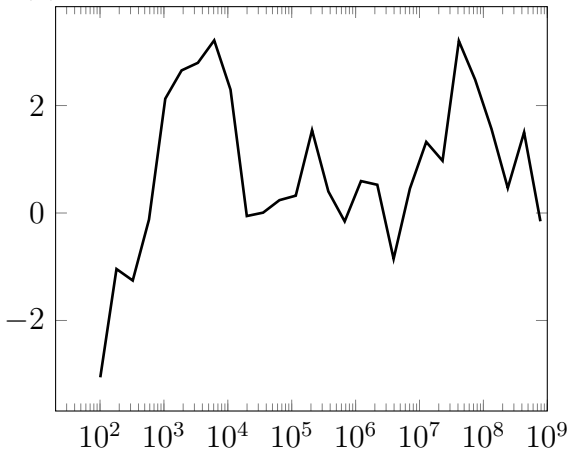
(a) Résultat de la première expérience.



(b) Résultat de la deuxième expérience.



(c) Résultat de la troisième expérience.



(d) Résultat de la quatrième expérience.

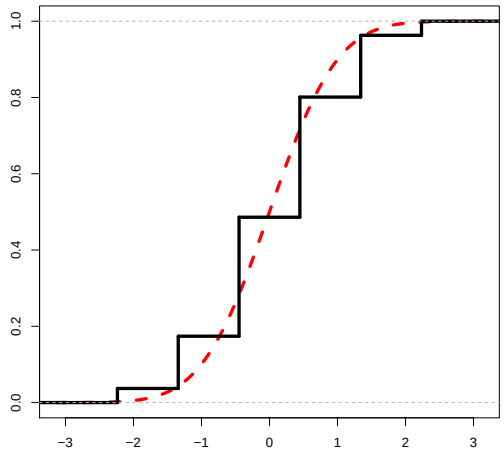
FIGURE 3.2 – La loi forte des grands nombres illustrée sur la figure précédente nous assure que $\bar{X}_n \to 1/2$, i.e., pour chaque expérience on a convergence de la moyenne empirique vers $1/2$. On s'intéresse maintenant à la vitesse à laquelle cette convergence a lieu et d'ainsi mieux comprendre pourquoi $\bar{X}_n$ n'a pas encore convergé sur la figure 3.1b. Les quatre figures représentent l'évolution, sur une échelle logarithmique en n, de la suite (Y_n) avec

$$Y_n = \left(\frac{n}{4\ln(\ln(n))}\right)^{1/2} \left(\bar{X}_n - 1/2\right).$$

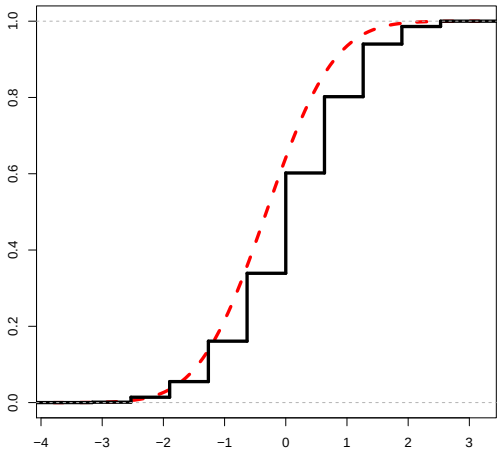
Ces résultats suggèrent que cette suite fluctue, ce qui est en fait une manifestation de la loi du logarithme itéré qui nous assure que, pour chaque expérience aléatoire, on a

$$\limsup_{n\to\infty} Y_n = 2 \quad \text{et} \quad \liminf_{n\to\infty} Y_n = -2.$$

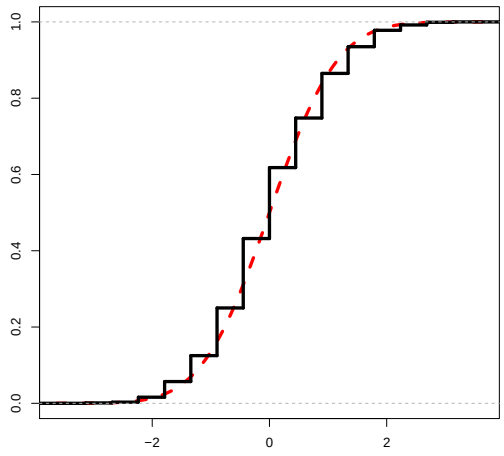
En particulier, il n'existe pas de développement asymptotique du genre $\bar{X}_n = 1/2 + \varepsilon_n + o(\varepsilon_n)$ avec une suite déterministe $\varepsilon_n \to 0$.



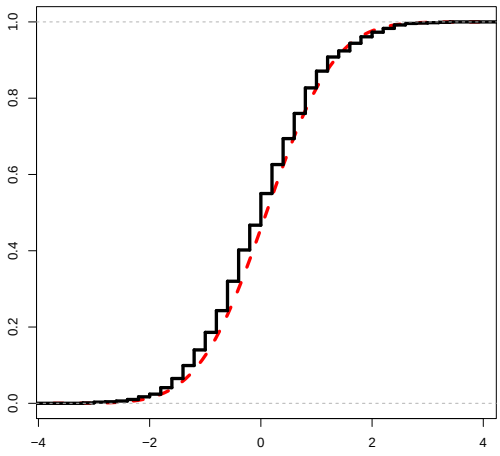
(a) Fonction de répartition de Z_5.



(b) Fonction de répartition de Z_{10}.



(c) Fonction de répartition de Z_{20}.



(d) Fonction de répartition de Z_{100}.

FIGURE 3.3 – Dans notre recherche de la vitesse de convergence de Z_n, nous avons été trop gourmands en cherchant $\varepsilon_n \to 0$ tel que $\bar{X}_n = 1/2 + \varepsilon_n + o(\varepsilon_n)$: une telle suite n'existe pas. Après tout, ce n'est peut-être pas surprenant vu que, dans un contexte probabiliste, il est peut-être plus raisonnable de n'exiger que des garanties concernant la probabilité que $u_n(\bar{X}_n - 1/2)$ appartienne à un certain intervalle $[a, b] \subset \mathbb{R}$. Si on définit $Z_n = u_n(\bar{X}_n - 1/2)$, on a

$$\mathbb{P}\left(a \leq u_n(\bar{X}_n - 1/2) \leq b\right) = F_{Z_n}(b) - F_{Z_n}(a)$$

et il devient naturel de s'intéresser au comportement asymptotique de la fonction de répartition de Z_n, illustré par les quatre figures ci-dessus : on voit clairement la convergence de F_{Z_n} vers une fonction limite (tracée en rouge), ce qui est une manifestation du théorème central limite qui garantit que si l'on choisit $u_n = n^{1/2}/2$, alors pour tout $x \in \mathbb{R}$ on a

$$F_{Z_n}(x) \underset{n\to\infty}{\longrightarrow} \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{x} e^{-y^2/2} \mathrm{d}y.$$

3.2 Convergence presque sûre et loi forte des grands nombres

3.2.1 Définition et loi forte des grands nombres

La convergence presque sûre est le mode de convergence dans la loi des grands nombres illustrée sur la Figure 3.1.

Définition 3.2.1. Soit X et $(X_n, n \in \mathbb{N})$ des variables aléatoires définies sur le même espace de probabilité. On dit que X_n converge **presque sûrement vers** X, ce que l'on note $X_n \overset{\text{p.s.}}{\to} X$, si et seulement si $X_n \to X$ avec probabilité un, i.e.,

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

Ainsi, $X_n \overset{\text{p.s.}}{\to} X$ si la convergence a lieu **pour** (presque toute) **réalisation de l'expérience aléatoire**

Théorème 3.2.1 (Loi forte des grands nombres)**.** *Si les $(X_n, n \geq 1)$ à valeurs dans $\mathbb{R}^m$ sont i.i.d. et intégrables, alors*

$$\frac{1}{n}(X_1 + \cdots + X_n) \overset{\text{p.s.}}{\longrightarrow} \mathbb{E}(X_1).$$

Nous ne prouverons pas ce résultat en toute généralité, mais proposerons une preuve plus loin basée sur le lemme de Borel–Cantelli. En outre, si l'on revient à l'exemple de la série de pile ou face on remarque qu'il existe des évènements ω tels que $\bar{X}_n(\omega)$ ne converge pas vers $1/2$: par exemple, l'évènement $\omega = \{0, 0, 0, \ldots\}$ qui correspond à obtenir une série infinie de face. Néanmoins, la loi forte des grands nombres nous dit précisément que la probabilité d'un tel évènement est nulle !

Proposition 3.2.2. *Si $X_n \overset{\text{p.s.}}{\to} X$ et f est continue, alors $f(X_n) \overset{\text{p.s.}}{\to} f(X)$.*

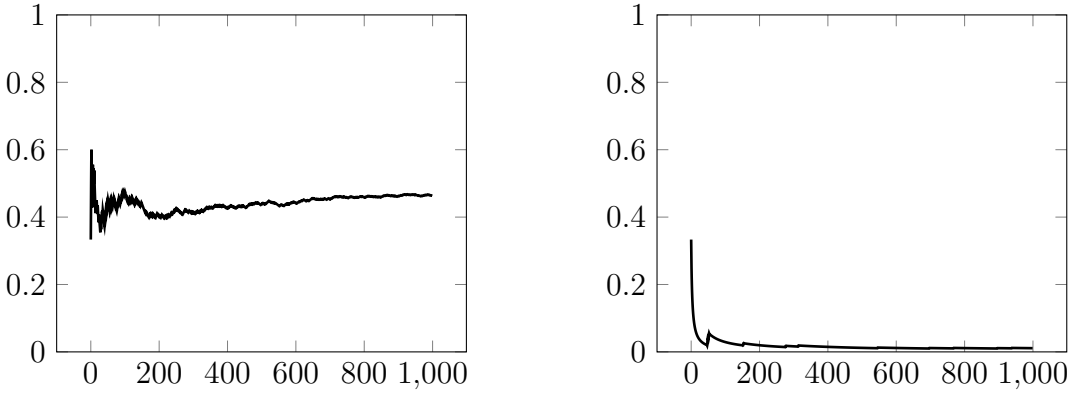
Démonstration. Pour ω tel que $X_n(\omega) \to X(\omega)$, on a par continuité $f(X_n(\omega)) \to f(X(\omega))$: ainsi, $\{X_n \to X\} \subset \{f(X_n) \to f(X)\}$ et donc $1 = \mathbb{P}(X_n \to X) \leq \mathbb{P}(f(X_n) \to f(X))$. ■

3.2.2 Convergence vers une variable aléatoire : l'exemple de l'urne de Pólya

La loi forte des grands nombres est l'archétype de la convergence presque sûre. Néanmoins, la limite est déterministe ce qui n'est pas forcément le cas en général. Nous présentons ici un exemple simple de suite de variables aléatoires qui converge presque sûrement vers une limite aléatoire : il s'agit de l'urne de Pólya.

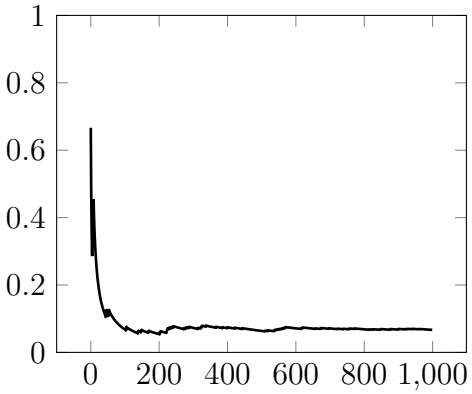
L'expérience est très simple : on a une urne avec des boules de deux couleurs, rouge et noir. Itérativement, on effectue l'opération suivante :

1. Tirer une boule de l'urne choisie uniformément au hasard ;
2. Remettre la boule dans l'urne **plus** une boule de la même couleur :
   - si la boule tirée est rouge, on remet donc la boule rouge **PLUS** une autre boule rouge ;
   - si la boule tirée est noire, on remet donc la boule noire **PLUS** une autre boule noire.

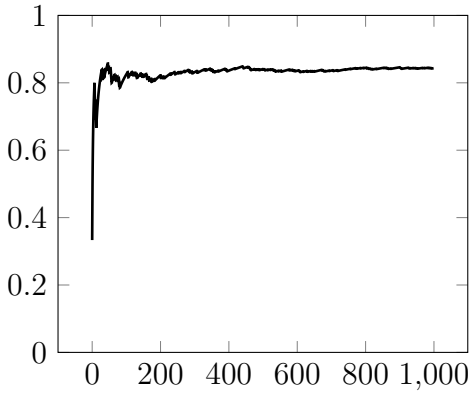


(a) ω avec $R_n(\omega) \to 0{,}46$.

(b) ω avec $R_n(\omega) \to 0{,}01$.



(c) ω avec $R_n(\omega) \to 0{,}06$.



(d) ω avec $R_n(\omega) \to 0{,}84$.

FIGURE 3.4 – Quatre réalisations d'une urne de Pólya : pour chaque réalisation, on s'intéresse à l'évolution de la fraction de boules rouges dans l'urne. On voit que cette fraction converge, mais la fraction asymptotique dépend de la réalisation. Il s'agit d'un cas typique de convergence presque sûre vers une limite aléatoire. En fait, on pourrait prouver que $R_n \overset{\text{p.s.}}{\to} R_\infty$ où R_∞ suit la loi uniforme sur $[0, 1]$.

Ainsi, si initialement on a 2 boules, après la première itération on aura 3 boules dans l'urne, puis 4, puis 5, etc. La Figure 3.4 présente quatre réalisations de cette expérience aléatoire, où l'on s'intéresse à la fraction R_n de boules rouges juste après la n-ième itération. Pour chaque expérience, on commence initialement avec une boule rouge et une boule noire (et donc $R_0 = 1/2$).

On voit sur cette figure que la suite R_n converge presque sûrement (i.e., pour chaque réalisation de l'expérience aléatoire), mais que la limite dépend de l'expérience et est donc aléatoire. Il y a une explication intuitive à ce comportement dû à un phénomène de renforcement. S'il est possible d'observer des fluctuations de la proportion de boules rouges au début (après tout, tant qu'il y a peu de boules tout peut arriver), une fois qu'il y a beaucoup de boules dans l'urne, il devient alors de plus en plus dur, et finalement impossible, d'inverser la situation et la proportion de boules rouges reste alors constante. On observe en effet que la convergence est très rapide, i.e., la valeur asymptotique est essentiellement déterminée au début de l'expérience et après quelques itérations il ne se passe plus grand chose.

3.2.3 Théorème de convergence dominée

Il est très fréquent que l'on ait besoin de la convergence de la moyenne d'une suite de variables aléatoires. Cela ne découle pas forcément de la convergence presque sûre, néanmoins si $X_n \overset{\text{p.s.}}{\to} X$ et les X_n sont bornés par une constante déterministe, alors on obtient la convergence des moyennes. Nous aurons plusieurs occasions dans les preuves suivantes de voir l'utilité d'un tel résultat.

Théorème 3.2.3 (Théorème de convergence dominée)**.** *Si $X_n \overset{\text{p.s.}}{\to} X$ et $\mathbb{P}(|X_n| \leq K) = 1$ pour un certain $K \in \mathbb{R}_+$, alors X_n et X sont intégrables et $\mathbb{E}(X_n) \to \mathbb{E}(X)$.*

3.2.4 Lemme de Borel–Cantelli

Prouver un résultat de convergence presque sûre peut s'avérer très délicat : par exemple, dans le cas de l'urne de Pólya discuté ci-dessus il n'existe même pas d'expression analytique de la limite R_∞. Néanmoins, il existe un critère de convergence presque sûre qui, lorsqu'il s'applique, est très efficace : il s'agit du lemme de Borel–Cantelli.

L'idée est simple : considérons pour $\varepsilon > 0$ la variable aléatoire

$$N(\varepsilon) = \sum_{k \geq 1} \mathbb{1}\left\{|X_k - X| \geq \varepsilon\right\}.$$

Ainsi, $N(\varepsilon)$ est le nombre de fois où X_k est à distance $\geq \varepsilon$ de X. Par définition de la convergence, on a alors $X_n(\omega) \to X(\omega)$ si et seulement si $N(\varepsilon)(\omega)$ est finie pour tout $\varepsilon > 0$. L'idée du lemme de Borel–Cantelli est très simple : si $\mathbb{E}(N(\varepsilon))$ est fini, alors $N(\varepsilon)$ est presque sûrement finie : puisque

$$\mathbb{E}(N(\varepsilon)) = \sum_{k \geq 1} \mathbb{P}\left(|X_k - X| \geq \varepsilon\right)$$

on obtient donc le critère suivant.

Théorème 3.2.4 (Lemme de Borel–Cantelli)**.** *Si pour tout $\varepsilon > 0$ on a*

$$\sum_{n \in \mathbb{N}} \mathbb{P}\left(|X_n - X| \geq \varepsilon\right) < \infty,$$

alors $X_n \overset{\text{p.s.}}{\to} X$.

Démonstration. De par la discussion ci-dessus, on voit que l'hypothèse du lemme de Borel–Cantelli implique que $N(\varepsilon)$ est presque sûrement fini pour tout $\varepsilon > 0$. Puisque l'union dénombrable d'évènements de probabilité nulle reste de probabilité nulle par (1.1), il s'ensuit que

$$\mathbb{P}\left(N(1/k) < \infty, \forall k \geq 1\right) = 1.$$

Puisque les événements $\{X_n \to X\}$ et $\bigcap_{k \geq 1}\{N(1/k) < \infty\}$ sont égaux, cela prouve le résultat. ∎

Nous appliquons maintenant ce lemme pour prouver la loi forte des grands nombres dans le cas particulier où $\mathbb{E}(e^{\theta X_1}) < \infty$ pour tout $\theta \in \mathbb{R}$.

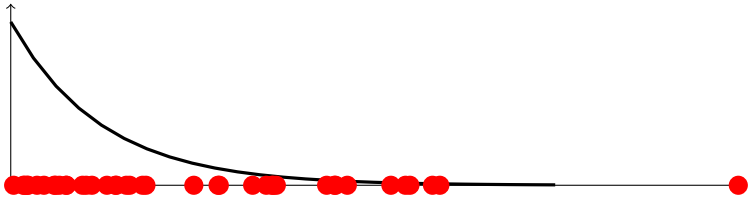


FIGURE 3.5 – Illustration de l'interprétation empirique de la densité discutée dans la Section 3.2.5.

Démonstration du Théorème 3.2.1 quand $\mathbb{E}(e^{\theta X_1}) < \infty$ *pour tout* $\theta \in \mathbb{R}$. Sans perte de généralité on supposera que $\mathbb{E}(X_1) = 0$. Soit $S_n = X_1 + \cdots + X_n$: alors

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_k\right| \geq \varepsilon\right) = \mathbb{P}\left(|S_n| \geq n\varepsilon\right) = \mathbb{P}\left(S_n \geq n\varepsilon\right) + \mathbb{P}\left(S_n \leq -n\varepsilon\right).$$

Pour $\theta \geq 0$, la fonction $x \mapsto e^{\theta x}$ est croissante et donc

$$\mathbb{P}\left(S_n \geq n\varepsilon\right) = \mathbb{P}\left(e^{\theta S_n - n\varepsilon} \geq 1\right) \leq e^{-\theta n\varepsilon}\mathbb{E}\left(e^{\theta S_n}\right)$$

où l'on a utilisé l'inégalité de Markov (Théorème 1.3.11) pour obtenir l'inégalité. Puisque les X_n sont i.i.d., on a $\mathbb{E}\left(e^{\theta S_n}\right) = \left[\mathbb{E}\left(e^{\theta X_1}\right)\right]^n$ et donc

$$\mathbb{P}\left(S_n \geq n\varepsilon\right) \leq \exp\left(-n(\theta\varepsilon - \varphi(\theta))\right) \quad \text{avec} \quad \varphi(\theta) = \ln \mathbb{E}(e^{\theta X_1}).$$

On admettra que φ est continûment dérivable de dérivée $\varphi'(\theta) = \mathbb{E}(X_1 e^{\theta X_1})/\mathbb{E}(e^{\theta X_1})$, ce que l'on pourrait prouver à l'aide du théorème de convergence dominée. Il s'ensuit que $\varphi(\theta)/\theta \to \varphi'(0) = 0$ lorsque $\theta \to 0$ et donc il existe $\theta_0 > 0$ tel que $\varphi(\theta_0) \leq \theta_0\varepsilon/2$. Considérant l'inégalité précédente pour ce θ_0, on obtient alors

$$\mathbb{P}\left(S_n \geq n\varepsilon\right) \leq \exp\left(-n(\theta_0\varepsilon - \varphi(\theta_0))\right) \leq \exp\left(-\frac{1}{2}n\varepsilon\theta_0\right)$$

ce qui prouve que $\sum_{n\geq 1} \mathbb{P}(S_n \geq n\varepsilon) < \infty$. De manière symétrique on prouve que la série de terme général $\mathbb{P}(S_n \leq -n\varepsilon)$ est sommable ce qui donne le résultat. ■

3.2.5 Interprétation empirique de la densité

La loi forte des grands nombres permet de porter un regard différent sur la densité. La Figure 3.5 représente par des points sur l'axe des abscisses un échantillon de 40 variables exponentielles i.i.d., notées $X_1, \ldots, X_{40}$. On remarque une "densité" de points décroissante : plus on s'éloigne de l'origine et plus les points sont loin les uns des autres.

En fait, cela reflète que la densité de la variable exponentielle, donnée par $x \in \mathbb{R}_+ \mapsto e^{-x}$, est décroissante. En effet, la loi des grands nombres nous assure que pour tout sous-ensemble $A \subset \mathbb{R}$, la densité empirique

$$\frac{1}{n}\sum_{k=1}^{n} \mathbb{1}\left\{X_k \in A\right\}$$

converge presque sûrement vers

$$\mathbb{E}(\mathbb{1}\left\{X_1 \in A\right\}) = \mathbb{P}(X_1 \in A) = \int_A f_{X_1}.$$

Ainsi, de manière empirique, le nombre de points qui tombent dans un intervalle donné est proportionnel à la densité : dans les zones où la densité est élevée, il y aura donc une forte accumulation de points.

3.3 Convergence en probabilité et loi faible des grands nombres

Définition 3.3.1. Soit X et $(X_n, n \in \mathbb{N})$ des variables aléatoires définies sur le même espace de probabilité. On dit que X_n converge **en probabilité vers** X, ce que l'on note $X_n \xrightarrow{\mathbb{P}} X$, si pour tout $\varepsilon > 0$ on a

$$\mathbb{P}\left(|X_n - X| \geq \varepsilon\right) \underset{n\to\infty}{\longrightarrow} 0.$$

Au vu de cette définition, la proposition suivante découle directement de l'inégalité de Bienaymé–Tchebycheff (Théorème 1.3.12).

Proposition 3.3.1. *Si* $\mathbb{E}(X_n) \to x$ *et* $\mathbb{V}\mathrm{ar}(X_n) \to 0$, *alors* $X_n \xrightarrow{\mathbb{P}} x$.

On note que convergence presque sûre implique convergence en probabilités : en effet, si $X_n \xrightarrow{\text{p.s.}} X$ alors pour tout $\varepsilon > 0$ on a $\mathbb{1}\{|X_n - X| \geq \varepsilon\} \xrightarrow{\text{p.s.}} 0$ et la convergence $X_n \xrightarrow{\mathbb{P}} X$ s'ensuit donc du théorème de convergence dominée.

Proposition 3.3.2. *Convergence presque sûre implique convergence en probabilités : si* $X_n \xrightarrow{\text{p.s.}} X$ *alors* $X_n \xrightarrow{\mathbb{P}} X$.

Par contre, la réciproque n'est pas vraie : un contre-exemple est donnée par la suite $X_n = B_n$ où les B_n sont des variables de Bernoulli indépendantes avec $\mathbb{E}(B_n) = 1/n$. Alors

$$\mathbb{P}(X_n \geq \varepsilon) = \mathbb{P}(B_n = 1) = \frac{1}{n}$$

et donc $X_n \xrightarrow{\mathbb{P}} 0$. Néanmoins, on peut prouver que

$$\mathbb{P}\left(\limsup_{n\to\infty} X_n = 1\right) = 1$$

et donc X_n ne converge pas presque sûrement vers 0.

Une conséquence directe du résultat précédent est la loi faible des grands nombres.

Théorème 3.3.3 (Loi faible des grands nombres)**.** *Si les* $(X_n, n \geq 1)$ *sont i.i.d. et intégrables, alors*

$$\frac{1}{n}\left(X_1 + \cdots + X_n\right) \xrightarrow{\mathbb{P}} \mathbb{E}(X_1).$$

Démonstration dans le cas de la variance finie. On fournit la preuve dans le cas de la variance finie où X_1 est de carré intégrable. On suppose sans perte de généralité que $\mathbb{E}(X_1) = 0$: l'inégalité de Bienaymé–Tchebycheff donne

$$\mathbb{P}\left(\frac{1}{n}\left|X_1 + \cdots + X_n\right| \geq \varepsilon\right) \leq \frac{1}{n^2\varepsilon^2}\mathbb{V}\mathrm{ar}(X_1 + \cdots + X_n) = \frac{\mathbb{V}\mathrm{ar}(X_1)}{n\varepsilon^2}$$

et il suffit donc de faire tendre $n \to \infty$. ■

Comparé à la loi forte des grands nombres, on prouve beaucoup plus facilement le résultat sous des hypothèses plus générales (variance finie pour la loi faible, tous les moments exponentiels finis pour la loi forte) : cela correspond au fait que la convergence en probabilité est un mode de convergence moins fort que la convergence presque sûre.

3.4 Convergence en loi et théorème central limite

Les convergences presque sûre et en probabilité introduites ci-dessus sont des convergences trajectorielles, dans le sens où $X_n - X$ tend vers 0 en un certain sens. En particulier, toutes les variables aléatoires doivent vivre sur le même espace de probabilité afin de donner un sens à la différence $X_n - X$.

La notion de convergence en loi est différente : on veut simplement que **les lois de X_n convergent**. Puisque la loi d'une variable aléatoire est indépendante de l'espace de probabilités sous-jacent, on peut avoir convergence en loi de variables aléatoires vivant dans des espaces distincts. Dans le cadre de ce cours on se limitera à la convergence en loi pour des variables aléatoires à valeurs dans $\mathbb{R}^m$, ce qui permet de définir la convergence en loi par la convergence des fonctions caractéristiques.

Définition 3.4.1. Soit X et $(X_n, n \in \mathbb{N})$ des variables aléatoires à valeurs dans $\mathbb{R}^m$. On dit que X_n converge **en loi vers** X, ce que l'on note $X_n \xrightarrow{\mathrm{L}} X$, si pour tout $t \in \mathbb{R}^m$ on a $\varphi_{X_n}(t) \to \varphi_X(t)$.

Par définition, la convergence en loi est donc indépendante de la structure de dépendance entre les variables (X_n) :

La convergence en loi de la suite (X_n) ne contient aucune information sur la corrélation entre les X_n !

En effet, dans les deux exemples suivants on a $X_n \xrightarrow{\mathrm{L}} X_1$ alors que dans le premier cas, les X_n sont indépendantes et dans le deuxième, elles sont parfaitement corrélées.

Exemple 3.4.1. Soit X_n des variables de Bernoulli de paramètre 1/2. Alors $X_n \xrightarrow{\mathrm{L}} X_1$ dans les deux cas suivants :

- les (X_n) sont indépendantes ;
- on tire X_1 puis on prend $X_n = X_1$: alors les X_n sont parfaitement corrélées.

Cet exemple illustre aussi que l'on peut avoir convergence en loi sans avoir convergence presque sûre ni convergence en probabilités. Par contre, convergence en probabilités (et donc, convergence presque sûre) implique convergence en loi.

Proposition 3.4.1. *Si $X_n \xrightarrow{\mathbb{P}} X$ alors $X_n \xrightarrow{\mathrm{L}} X$.*

Démonstration. On a

$$|\varphi_{X_n}(t) - \varphi_X(t)| = \left|\mathbb{E}\left(e^{itX_n} - e^{itX}\right)\right| \leq \mathbb{E}\left|e^{itX_n} - e^{itX}\right| \leq 2\mathbb{E}\left[\min\left(1, t\left|X_n - X\right|\right)\right]$$

où l'on a utilisé l'inégalité $|e^{iz} - e^{iz'}| \leq 2\min(1, |z - z'|)$ pour tous $z, z' \in \mathbb{R}$. Pour $\varepsilon \in]0, 1[$, on a

$$\begin{aligned}\mathbb{E}\left[\min\left(1, t\left|X_n - X\right|\right)\right] = \mathbb{E}\left[\min\left(1, t\left|X_n - X\right|\right); t|X_n - X| \leq \varepsilon\right] \\ + \mathbb{E}\left[\min\left(1, t\left|X_n - X\right|\right); t|X_n - X| \geq \varepsilon\right]\end{aligned}$$

ce qui donne

$$\mathbb{E}\left[\min\left(1, t\left|X_n - X\right|\right)\right] \leq \varepsilon + \mathbb{P}\left(\left|X_n - X\right| \geq \varepsilon/t\right).$$

Puisque $X_n \xrightarrow{\mathbb{P}} X$, on obtient donc

$$\limsup_{n\to+\infty} \left|\varphi_{X_n}(t) - \varphi_X(t)\right| \leq \varepsilon.$$

Puisque cette borne est valable pour tout $\varepsilon \in]0,1[$, il ne reste plus qu'à faire tendre $\varepsilon \to 0$ pour obtenir le résultat. ■

Puisque convergence presque sûre implique convergence en probabilité, on a donc

$$\boxed{\left(X_n \xrightarrow{\text{p.s.}} X\right) \Longrightarrow \left(X_n \xrightarrow{\mathbb{P}} X\right) \Longrightarrow \left(X_n \xrightarrow{\text{L}} X\right).}$$

On énonce maintenant plusieurs critères de convergence en loi, qui font notamment intervenir les transformées de la Section 1.5.

Proposition 3.4.2. *Chacune des conditions suivantes est une condition nécessaire et suffisante pour que $X_n \xrightarrow{\text{L}} X$:*

- *X_n et X sont à valeurs dans $\mathbb{N}^m$ et $\mathbb{P}(X_n = x) \to \mathbb{P}(X = x)$ pour tout $x \in \mathbb{N}^m$;*
- *X_n et X sont à valeurs dans $\mathbb{N}^m$ et $\phi_{X_n}(z) \to \phi_X(z)$ pour tout $z \in [-1,1]^m$;*
- *X_n et X sont à valeurs dans $\mathbb{R}_+^m$ et $L_{X_n}(\lambda) \to L_X(\lambda)$ pour tout $\lambda \in \mathbb{R}_+^m$;*
- *X_n et X sont à valeurs dans $\mathbb{R}^m$, X est absolument continue et $F_{X_n}(x) \to F_X(x)$ pour tout $x \in \mathbb{R}_+^m$;*
- *X_n et X sont à valeurs dans $\mathbb{R}^m$ et $F_{X_n}(x) \to F_X(x)$ pour tout $x \in \mathbb{R}_+^m$ tel que $\mathbb{P}(X = x) = 0$.*

L'exemple ci-dessous montre que l'on peut avoir $X_n \xrightarrow{\text{L}} X$ mais que $F_{X_n}(x) \not\to F_X(x)$ si $\mathbb{P}(X = x) \neq 0$.

Exemple 3.4.2. Si $X_n = 1/n$, alors $X_n \xrightarrow{\text{L}} 0$ (puisque $X_n \xrightarrow{\text{p.s.}} 0$) mais $F_{X_n}(0) = \mathbb{P}(X_n \leq 0) = 0$ ne tend pas vers $F_X(0) = 1$.

Les résultats suivants sont très utiles.

Proposition 3.4.3. *Si $X_n \xrightarrow{\text{L}} X$, alors $f(X_n) \xrightarrow{\text{L}} f(X)$ pour toute fonction continue f.*

Proposition 3.4.4 (Lemme de Slutsky)**.** *Si $X_n \xrightarrow{\text{L}} X$ et $Y_n \xrightarrow{\text{L}} c$ avec c une constante, alors $(X_n, Y_n) \xrightarrow{\text{L}} (X, c)$.*

Avant d'énoncer le théorème central limite, qui est le deuxième résultat le plus important de la théorie des probabilités après la loi des grands nombres, on montre deux exemples simples de convergence en loi qui justifient de voir la loi de Poisson comme la loi des évènements rares ainsi que la loi exponentielle comme l'équivalent continu de la loi géométrique.

Proposition 3.4.5. *Soit $\lambda \in \mathbb{R}_+$ et X_n qui suit une loi binomiale de paramètre $(n, \lambda/n)$: alors $X_n \xrightarrow{\text{L}} X$ où X suit une loi de Poisson de paramètre λ.*

Démonstration. On peut écrire $X_n = I_n^1 + \cdots + I_n^n$ où les I_n^k sont i.i.d. et suivent une loi de Bernoulli de paramètre λ/n : il s'ensuit que

$$\phi_{X_n}(z) = \left[\mathbb{E}\left(z^{I_n^1}\right)\right]^n = \left(1 - \frac{\lambda}{n}(1-z)\right)^n$$

qui tend vers $e^{-\lambda(1-z)}$ quand $n \to \infty$. Puisqu'on reconnaît la fonction génératrice de la loi de Poisson de paramètre λ, cela prouve le résultat par la proposition précédente. ■

Proposition 3.4.6. *Soit $\lambda \in \mathbb{R}_+$ et X_n pour $n \geq \lambda$ qui suit une loi géométrique de paramètre λ/n : alors $X_n/n \xrightarrow{\mathrm{L}} X$ où X suit une loi exponentielle de paramètre λ.*

Démonstration. Pour $x \in \mathbb{R}_+$, on a

$$1 - F_{X_n/n}(x) = 1 - \mathbb{P}\left(X_n/n \leq x\right) = \mathbb{P}\left(X_n/n > x\right) = \mathbb{P}\left(X_n > nx\right) = \left(1 - \frac{\lambda}{n}\right)^{\lceil nx \rceil}$$

ce qui montre que $F_{X_n/n}(x) \to 1 - e^{-\lambda x}$. D'un autre côté, puisque la densité de la loi exponentielle de paramètre λ est $\lambda e^{-\lambda x}\mathbb{1}\{x \geq 0\}$, on a

$$F_X(x) = \mathbb{P}(X \leq x) = \int_0^x \lambda e^{-\lambda u} \mathrm{d}u = 1 - e^{-\lambda x}.$$

On a donc bien montré $F_{X_n}(x) \to F_X(x)$ pour tout $x \in \mathbb{R}$ ce qui montre que $X_n \xrightarrow{\mathrm{L}} X$ par la Proposition 3.4.2. ■

Théorème 3.4.7 (Théorème central limite)**.** *Soit $(X_n, n \geq 1)$ i.i.d. avec $\mathbb{E}(X_1) = 0$ et $\mathbb{V}\mathrm{ar}(X_1) = 1$: alors*

$$\frac{1}{n^{1/2}} \sum_{k=1}^n X_k \xrightarrow{\mathrm{L}} X$$

où X suit une loi standard normale.

Ebauche de démonstration. Puisque les X_k sont i.i.d., on a

$$\varphi_{n^{-1/2}(X_1+\cdots+X_n)}(t) = \mathbb{E}\left[\exp\left(\frac{it}{n^{1/2}}\sum_{k=1}^n X_k\right)\right] = \left[\varphi_{X_1}(tn^{-1/2})\right]^n.$$

Pour $\varepsilon \to 0$, le développement limité $e^\varepsilon = 1 + \varepsilon + \frac{\varepsilon^2}{2} + o(\varepsilon^2)$ suggère

$$\varphi_{X_1}(\varepsilon) = \mathbb{E}\left(e^{i\varepsilon X_1}\right) = \mathbb{E}\left(1 + i\varepsilon X_1 - \frac{\varepsilon^2 X_1^2}{2} + o(\varepsilon^2)\right)$$

et puisque $\mathbb{E}(X_1) = 0$ et $\mathbb{V}\mathrm{ar}(X_1) = 1$, on admettra que cela donne $\varphi_{X_1}(\varepsilon) = 1 - \varepsilon^2/2 + o(\varepsilon^2)$. Ainsi,

$$\mathbb{E}\left[\exp\left(\frac{it}{n^{1/2}}\sum_{k=1}^n X_k\right)\right] = \left[1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right]^n \underset{n\to\infty}{\longrightarrow} e^{-t^2/2}.$$

Puisque $e^{-t^2/2}$ est la fonction caractéristique de X, cela donne le résultat. ■

Puisque X est absolument continue, la Proposition 3.4.2 montre que

$$F_{n^{-1/2}(X_1+\cdots+X_n)}(x) \to F_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} \mathrm{d}y$$

ce qui était déjà illustré numériquement sur la Figure 3.3.

Si X est de variance finie, alors la variable aléatoire $(X - \mathbb{E}(X))/\sqrt{\mathbb{V}\mathrm{ar}(X)}$ est centrée (i.e., de moyenne nulle) et réduite (i.e., de variance 1) : ainsi, le théorème central limite s'applique aussi à une suite de variables i.i.d. dès lors que leur variance commune est finie.

Proposition 3.4.8. *Soit $(X_n, n \geq 1)$ i.i.d. avec $\mathbb{V}\mathrm{ar}(X_1) \in (0, \infty)$: alors*

$$\frac{1}{n^{1/2}\sqrt{\mathbb{V}\mathrm{ar}(X_1)}} \sum_{k=1}^{n} (X_k - \mathbb{E}(X_1)) \xrightarrow{\mathrm{L}} X$$

où X suit une loi standard normale.

3.5 Généralisation à la dimension ≥ 1

La convergence en probabilités n'a été définie qu'en dimension 1, et plusieurs des preuves ci-dessus ne considérent que ce cas. Néanmoins, tout se généralise en dimension $d \geq 1$ en remplaçant les valeurs absolues par n'importe quelle norme sur $\mathbb{R}^d$.

3.6 Fiche de synthèse

Convergence en loi : $X_n \xrightarrow{\text{L}} X$ si $F_{X_n}(x) \underset{n\to+\infty}{\to} F_X(x)$ en tout x où F_X est continue, ce qui équivaut à $\varphi_{X_n}(t) \underset{n\to+\infty}{\to} \varphi_X(t)$ pour tout t (rappel : $\varphi_X(t) = \mathbb{E}(e^{itX})$ est la fonction caractéristique de X).

Convergence en probabilité : $X_n \xrightarrow{\mathbb{P}} X$ si, pour tout $\varepsilon > 0$,

$$\lim_{n\to+\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

Convergence presque sûre : $X_n \xrightarrow{\text{p.s.}} X$ si $\mathbb{P}\left(\{\omega \in \Omega \ ; \ \lim_{n\to+\infty} X_n(\omega) = X(\omega)\}\right) = 1.$

Propriété : La convergence p.s. implique la convergence en probabilité, qui implique la convergence en loi. La réciproque est fausse en général, mais

convergence en loi vers une constante $\Leftrightarrow$ convergence en probabilité vers cette constante.

Loi forte des grands nombres : si les X_i sont des v.a. indépendantes de même loi, d'espérance m, alors $\dfrac{X_1 + \cdots + X_n}{n} \xrightarrow{p.s.} m$ (v.a. constante).

Théorème Central Limite : si les X_i sont de plus de variance σ^2 finie, alors

$$\frac{X_1 + \cdots + X_n - nm}{\sqrt{n} \times \sigma} \xrightarrow{L} Z$$

avec Z de loi normale $\mathcal{N}(0,1)$.
Conséquence : $\mathbb{P}_{X_1+\cdots+X_n} \approx \mathcal{N}(nm, n\sigma^2)$ ou encore, $X_1 + \cdots + X_n \approx nm + \sigma Z$ avec $Z \sim \mathcal{N}(0,1)$.

3.7 Exercices

Les exercices précédés d'une flèche ↪ sont des exercices d'application directs du cours.

↪ **Exercice 3.1**

1. Soit (U_n) i.i.d. uniformément réparties sur $[0,1]$ et $M_n = \max_{k=1,\dots,n} U_k$: montrez que la suite (M_n) converge presque sûrement. Quelle est sa limite ?

↪ **Exercice 3.2**

1. Montrez que si $X_n \xrightarrow{\mathrm{L}} c$ avec c une constante, alors $X_n \xrightarrow{\mathbb{P}} c$.

2. Montrez en utilisant les Propositions 3.4.3 et 3.4.4 que si $X_n \xrightarrow{\mathrm{L}} X$ et $X_n - Y_n \xrightarrow{\mathrm{L}} 0$, alors $Y_n \xrightarrow{\mathrm{L}} X$.

Exercice 3.3 *(Convergence jointe)*

1. Montrez que $\mathbb{P}(X+Y \geq \varepsilon) \leq \mathbb{P}(X \geq \varepsilon/2) + \mathbb{P}(Y \geq \varepsilon/2)$.

2. Montrez en utilisant la question précédente que si $X_n \xrightarrow{\mathbb{P}} 0$ et $Y_n \xrightarrow{\mathbb{P}} 0$, alors $X_n + Y_n \xrightarrow{\mathbb{P}} 0$. Déduisez-en que si $X_n \xrightarrow{\mathbb{P}} X$ et $Y_n \xrightarrow{\mathbb{P}} Y$, alors $(X_n, Y_n) \xrightarrow{\mathbb{P}} (X, Y)$.

3. Montrez que le résultat précédent n'est plus valable pour la convergence en loi, i.e., que l'on peut avoir la convergence en loi des marginales mais pas la convergence en loi de la loi jointe.

↪ **Exercice 3.4** *(Valeurs extrêmes)*
On considère $(X_k, k \in \mathbb{N}^*)$ une suite de variables aléatoires i.i.d. de loi commune la loi de X : on s'intéresse dans cette question à la convergence en loi du maximum $M_n = \max(X_1, \dots, X_n)$.

1. Montrez que $\mathbb{P}(M_n \leq x) = \mathbb{P}(X \leq x)^n$.

2. On suppose que X suit une loi uniforme sur $[a,b]$: quelle est la limite de M_n (on pourra utilise l'exercice 3.1) ? Montrez que $n(b - M_n)$ converge en en loi : quelle est sa limite ?

3. On suppose que X suit une loi Beta de paramère $(\alpha, 2)$, i.e., sa densité f est donnée par

$$f(x) = \alpha(\alpha+1)x^{\alpha-1}(1-x)\mathbb{1}\{x \in [0,1]\}.$$

En utilisant

$$\int_{1-\varepsilon}^{1} f(x)\mathrm{d}x \sim f'(1)\int_{1-\varepsilon}^{1}(x-1)\mathrm{d}x = \frac{\alpha(\alpha+1)}{2}\varepsilon^2,$$

montrez que $n^{1/2}(1 - M_n)$ converge en loi. Montrez que la variable aléatoire limite est absolument continue et donnez une expression de sa densité.

4. On suppose que X suit une loi exponentielle de paramètre λ. Montrez que $M_n - (1/\lambda)\log n$ converge en loi et exprimez la limite.

5. On suppose que X vérifie $\mathbb{P}(X \geq x) = (1+x)^{-\alpha}$: trouvez β tel que M_n/n^β converge en loi et exprimez la limite.

↪ **Exercice 3.5** *(Physique statistique)*
Soit X une variable aléatoire absolument continue.

1. Calculez la loi de $X_\varepsilon = \varepsilon\lfloor X/\varepsilon \rfloor$ pour $\varepsilon > 0$ et comparez sur un même graphique la loi de X et celle de X_ε.

2. Montrez que $X_\varepsilon \xrightarrow{\text{p.s.}} X$ lorsque $\varepsilon \downarrow 0$.

3. Pour $\varepsilon > 0$ on considère X_ε une variable aléatoire discrète de support $\{\varepsilon k : k \in \mathbb{Z}\}$ telle que

$$\mathbb{P}(X_\varepsilon = \varepsilon k) = \frac{1}{Z_\varepsilon} f_X(\varepsilon k),\ k \in \mathbb{Z},$$

avec Z_ε la constante de normalisation. Montrez que $X_\varepsilon \xrightarrow{\text{L}} X$.

Exercice 3.6 *(Lien avec l'analyse fonctionnelle)*
L'énoncé suivant est la question 2 de l'exercice 4 de la PC8&PC9 d'analyse fonctionnelle.

---

*Une suite de fonctions (U_n) est appelée **unité approchée pour la convolution** si elle vérifie les propriétés suivantes :*

1. $\forall n \in \mathbb{N}$, $\forall x \in \mathbb{R}$, $U_n(x) \geq 0$;
2. $\forall n \in \mathbb{N}$, $\int U_n(x)\mathrm{d}x = 1$;
3. $\forall \varepsilon > 0$, $\int_{|x|>\varepsilon} U_n(x)\mathrm{d}x \to 0$ *lorsque* $n \to \infty$.

- Montrez que la suite de fonctions $U_n(x) = \frac{n}{\sqrt{2\pi}} e^{-x^2 n^2/2}$ est une unité approchée.
- Montrez que pour tout fonction $f \in L^1(\mathbb{R})$, $f * U_n$ appartient à $L^1(\mathbb{R}) \cap \mathcal{C}^\infty(\mathbb{R})$.
- Soit f une fonction de $L^1(\mathbb{R}) \cap C^0(\mathbb{R})$. On suppose de plus que f est bornée sur $\mathbb{R}$. Montrez que, pour tout réel fixé x_0, $(f * U_n)(x_0)$ converge vers $f(x_0)$ (utiliser le théorème de convergence dominée).

---

1. Reformulez-le et résolvez-le en terme probabiliste.

Problème 3.7

Dans ce problème, on considère $(X_k, k \in \mathbb{N}^*)$ une suite de variables aléatoires indépendantes et l'on exhibe des cas où la suite converge en loi vers 0 mais pas presque sûrement. On regarde tout d'abord le cas de variables de Bernoulli puis le cas de variables exponentielles, pour lesquelles on exhibe des conditions nécessaires et suffisantes de convergence en loi et presque sûre vers 0.

Dans la première partie du problème, X_k est une variable aléatoire de Bernoulli de paramètre p_k.

1. A quelle condition nécessaire et suffisante a-t-on $X_n \xrightarrow{\text{L}} 0$?

Dans le reste de l'exercice on suppose que $p_k \to 0$ et on définit $N_n = \sum_{k=1}^n X_k$.

2. Montrez que la suite N_n converge presque sûrement vers une variable aléatoire N_∞ à valeurs dans $\mathbb{N} \cup \{\infty\}$.

3. Montrez que $X_n \xrightarrow{\text{p.s.}} 0$ si et seulement si $\mathbb{P}(N_\infty < \infty) = 1$.

4. Déduisez-en en considérant la moyenne de N_∞ que si $\sum_k p_k < \infty$, alors $X_n \xrightarrow{\text{p.s.}} 0$. De quel résultat du cours ce résultat est-il un cas particulier ?

Les questions 5 et 6 visent à prouver la réciproque, i.e., que si $\sum_k p_k = \infty$ alors X_n ne converge pas presque sûrement vers 0.

5. On suppose que pour tout $K \geq 0$, on a $\mathbb{P}(N_n \leq K) \to 0$: montrez que cela implique que X_n ne converge pas presque sûrement vers 0.

6. Montrez en utilisant l'inégalité de Markov pour la première inégalité que

$$\mathbb{P}\left(N_n \leq K\right) \leq e^K \mathbb{E}\left(e^{-N_n}\right) = e^K \exp\left(\sum_{k=1}^n \log\left(1 - p_k(1 - e^{-1})\right)\right)$$

et conclure.

7. La suite X_n converge-t-elle vers 0 pour $p_k = 1/k$?

8. Soit U uniforme sur $[0,1]$ et $X'_n = \mathbb{1}\{U \leq 1/n\}$: en quel sens la suite X'_n converge-t-elle vers 0 ? Cela contredit-il le résultat de la question précédente ?

On suppose maintenant que X_k est une variable exponentielle de paramètre λ_k, et on définit $N_n(\varepsilon) = \sum_{k=1}^n \mathbb{1}\{X_k \geq \varepsilon\}$.

9. Montrez que pour tout $\varepsilon > 0$, la suite $N_n(\varepsilon)$ converge presque sûrement : on notera $N_\infty(\varepsilon)$ la limite.

10. Montrez que

$$X_n \xrightarrow{\text{p.s.}} 0 \iff \forall k \in \mathbb{N}^*,\ \mathbb{P}(N_\infty(1/k) < \infty) = 1.$$

11. En déduire à l'aide du cas Bernoulli que

$$X_n \overset{\text{p.s.}}{\to} 0 \iff \forall \varepsilon > 0, \sum_{k \geq 1} e^{-\lambda_k \varepsilon} < \infty.$$

12. Trouvez λ_k tel que la suite X_n converge en loi mais pas presque sûrement vers 0.

Chapitre 4

Vecteurs gaussiens

Les vecteurs gaussiens généralisent en dimension ≥ 1 la loi normale sur $\mathbb{R}$ introduite dans la Section 2.5.3 et qui est apparue comme limite universelle dans le théorème central limite sur $\mathbb{R}$ (Théorème 3.4.7). Ils jouent un rôle prépondérant dans de nombreux domaines d'application de la théorie des probabilités, et notamment en statistique inférentielle et en traitement du signal, mais aussi en optimisation où les processus gaussiens sont utilisés dans la construction de méta-modèles.

4.1 Définition et propriétés élémentaires

4.1.1 Vecteur gaussien standard

Nous définissons ci-dessous un vecteur gaussien comme étant l'image par une application affine d'un vecteur gaussien standard, puis montrons dans le Théorème 4.4.1 que cette définition est équivalente à la définition classique d'un vecteur gaussien comme vecteur dont toute combinaison linéaire des coordonnées suit une loi normale sur $\mathbb{R}$.

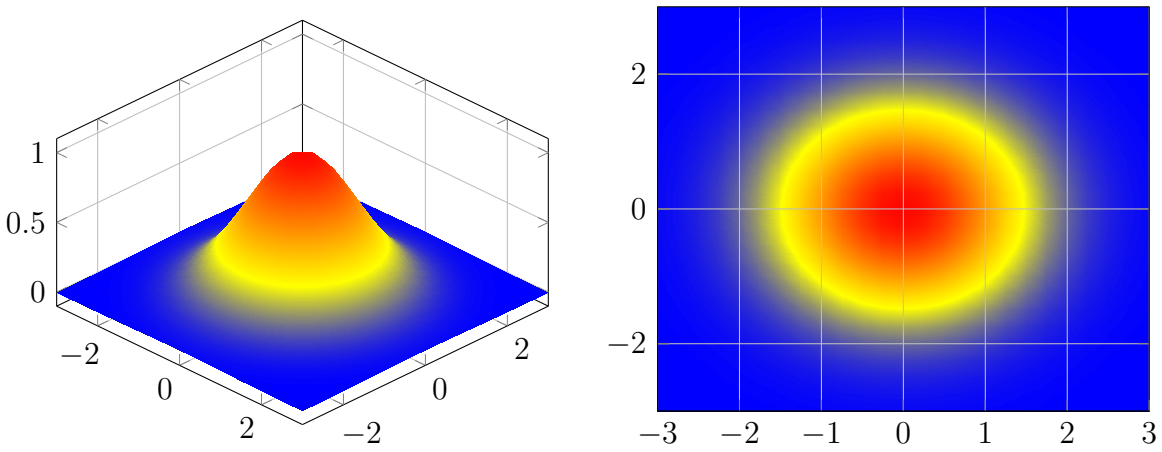
Définition 4.1.1 (Loi normale standard sur $\mathbb{R}^n$, vecteur gaussien standard)**.** La loi normale standard sur $\mathbb{R}^n$ est la loi du vecteur $(X_1, \ldots, X_n)$ où les X_i sont i.i.d. et suivent une loi standard normale sur $\mathbb{R}$. On dit alors que le vecteur $(X_1, \ldots, X_n)$ est un vecteur gaussien standard.

En particulier, si X est un vecteur gaussien standard, alors $\mathbb{E}(X) = 0$ et $\mathbb{V}\mathrm{ar}(X) = I_n$, l'identité de $\mathbb{R}^{n\times n}$. Par ailleurs, puisque la densité de variables indépendantes est égale au produit des densités (Théorème 2.6.5), et que la densité de la loi normale standard sur $\mathbb{R}$ est $(2\pi)^{-1/2}e^{-x^2/2}$, il s'ensuit directement que la densité de X est donnée par

$$f_X(x) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-x_k^2/2} = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}x^T x\right), \ x = (x_1, \ldots, x_n) \in \mathbb{R}^n. \tag{4.1}$$

Cette densité est représentée sur la Figure 4.1a et la Figure 4.1b illustre les courbes d'iso-densité $\{x : f(x) = c\}$ qui correspondent à des cercles $\{x : x^T x = c'\}$. On note aussi que la fonction caractéristique d'un vecteur gaussien standard X est donnée par

$$\varphi_X(t) = \exp\left(-\frac{1}{2}t^T t\right), \ t \in \mathbb{R}^n. \tag{4.2}$$



(a) Vue de profil. (b) Vue de haut.

FIGURE 4.1 – Densité de la loi normale standard sur $\mathbb{R}^2$ vue de deux angles différents. La Figure 4.1b illustre que les courbes iso-densité sont des cercles.

En effet,

$$\begin{aligned}
\varphi_X(t) &= \mathbb{E}(e^{it^T X}) && \text{(définition de } \varphi_X\text{)} \\
&= \mathbb{E}(e^{i(t_1X_1+\cdots+t_nX_n)}) \\
&= \prod_{k=1}^{n} \mathbb{E}(e^{it_kX_k}) && \text{(indépendance des } X_k\text{)} \\
&= \prod_{k=1}^{n} e^{-t_k^2/2} && \text{(Proposition 2.5.6)}
\end{aligned}$$

ce qui donne bien le résultat annoncé. On remarquera que, à un coefficient multiplicatif près, la densité et sa fonction caractéristique sont les mêmes, i.e., la transformée de Fourier laisse la loi normale invariante.

4.1.2 Vecteur gaussien

On définit maintenant un vecteur gaussien comme l'image par une application affine d'un vecteur gaussien standard.

Définition 4.1.2 (Loi normale sur $\mathbb{R}^n$, vecteur gaussien)**.** Une variable aléatoire $X \in \mathbb{R}^n$ est un vecteur gaussien s'il existe $\mu \in \mathbb{R}^n$, $m \in \mathbb{N}^*$, $M \in \mathbb{R}^{n\times m}$ et Y qui suit une loi normale standard sur $\mathbb{R}^m$ tels que $X = MY + \mu$. La loi d'un vecteur gaussien est appelée loi normale ou loi gaussienne.

Il découle immédiatement de cette définition que si X est un vecteur gaussien, alors **toute transformation affine de X reste un vecteur gaussien**. En particulier, si $X = (X_1, \ldots, X_n)$, alors chaque X_k est un vecteur gaussien. Néanmoins,

La réciproque n'est pas vraie en général !

En effet, on peut avoir X_1 et X_2 qui suivent une loi normale mais (X_1, X_2) n'est pas un vecteur gaussien comme le montre l'exemple ci-dessous. En revanche, nous verrons dans la Proposition 4.1.2 que la réciproque est vraie sous condition d'indépendance.

Exemple 4.1.1. Soit X_1 qui suit une loi normale standard, ε à valeurs dans $\{1,-1\}$ indépendante de X_1 avec $\mathbb{P}(\varepsilon=1)=\mathbb{P}(\varepsilon=-1)=1/2$ et $X_2=\varepsilon X_1$: alors X_2 suit aussi une loi normale standard et la propriété $\mathbb{P}(X_1+X_2=0)=1/2$ empêche X_1+X_2 d'être absolument continue et a fortiori de suivre une loi normale. Or, si (X_1,X_2) est un vecteur gaussien, il suit directement de la définition que toute combinaison linéaire de X_1 et X_2 devrait suivre une loi normale : puisque X_1+X_2 ne suit pas une loi normale, (X_1,X_2) ne peut pas être un vecteur gaussien.

Dans la Définition 4.1.2, il y a unicité du vecteur μ puisque par linéarité de l'espérance on a $\mu=\mathbb{E}(X)$:

$$\mathbb{E}(X)=M\mathbb{E}(Y)+\mu=\mu.$$

La matrice M n'est quant à elle pas forcément unique (cela sera discuté plus en détail en Section 4.3) mais par contre, elle satisfait nécessairement $MM^T=\mathbb{V}\mathrm{ar}(X)$. En effet, on a

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(X) &= \mathbb{V}\mathrm{ar}(MY+\mu) && \text{(définition de } X\text{)}\\
&= \mathbb{V}\mathrm{ar}(MY) && \text{(invariance de la variance par translation)}\\
&= \mathbb{E}(MY(MY)^T) && \text{(définition de la variance)}\\
&= \mathbb{E}(MYY^TM^T) && \text{(algèbre linéaire)}\\
&= M\mathbb{E}(YY^T)M^T && \text{(linéarité de l'espérance)}\\
&= M\mathbb{V}\mathrm{ar}(Y)M^T && \text{(définition de la variance)}
\end{aligned}$$

qui vaut bien MM^T puisque, comme mentionné précédemment, la matrice de covariance d'un vecteur gaussien standard est égale à l'identité. Cette remarque permet de calculer la fonction caractéristique d'un vecteur gaussien : en effet, pour $t\in\mathbb{R}^n$ on a

$$\varphi_X(t)=\mathbb{E}(e^{it^TX})=\mathbb{E}(e^{it^TMY+it^T\mu})=e^{it^T\mu}\mathbb{E}(e^{it^TMY})=e^{it^T\mu}\varphi_Y(M^Tt)$$

et donc en utilisant la formule 4.2 et le fait que $M^TM=\mathbb{V}\mathrm{ar}(X)$ et que $\mu=\mathbb{E}(X)$, on obtient finalement le résultat suivant.

Proposition 4.1.1. *Si X est un vecteur gaussien, alors il est de carré intégrable et sa fonction caractéristique est donnée par*

$$\varphi_X(t)=\exp\left(-\frac{1}{2}t^T\mathbb{V}\mathrm{ar}(X)t+it^T\mathbb{E}(X)\right),\ t\in\mathbb{R}^n.$$

Une conséquence très importante de ce résultat est que

La loi d'un vecteur gaussien est uniquement déterminée par son espérance et sa variance.

En particulier, toute propriété concernant la loi d'un vecteur gaussien est "visible" sur son espérance et sa variance. On peut ainsi facilement prouver les résultats suivants. On rappelle que le second point n'est pas vrai sans l'hypothèse que (X_1,X_2) est un vecteur gaussien, cf. la discussion à la fin de la Section 1.4.4.

Proposition 4.1.2. *Soit X_1 et X_2 deux vecteurs gaussiens. Alors :*

- *si X_1 et X_2 sont indépendantes, alors (X_1,X_2) est un vecteur gaussien ;*

- *si X_1 et X_2 sont décorrélées, i.e., $\mathbb{C}\mathrm{ov}(X_1, X_2) = 0$, **et que** (X_1, X_2) est un vecteur gaussien, alors X_1 et X_2 sont indépendantes.*

Démonstration. Supposons que X_1 et X_2 sont indépendantes et définissons $X = (X_1, X_2)$. Puisque la fonction caractéristique caractérise la loi, pour montrer le résultat il suffit de montrer que $\varphi_X(t) = e^{-\frac{1}{1}t^T\mathbb{V}\mathrm{ar}(X)t + it^T\mathbb{E}(X)}$ pour tout $t \in \mathbb{R}^n$, avec $n = n_1 + n_2$ où n_k est la dimension ambiante de X_k. On fixe donc $t \in \mathbb{R}^n$, que l'on écrit $t = (t_1, t_2)$ avec $t_k \in \mathbb{R}^{n_k}$. Puisque X_1 et X_2 sont indépendantes, on a

$$\varphi_X(t) = \varphi_{X_1,X_2}(t_1, t_2) = \varphi_{X_1}(t_1)\varphi_{X_2}(t_2)$$

et la Proposition 4.1.1 donne donc

$$\varphi_{X_1,X_2}(t_1, t_2) = \exp\left(-\frac{1}{2}\sum_{k=1,2} t_k^T\mathbb{V}\mathrm{ar}(X_k)t_k + i\sum_{k=1,2} t_k^T\mathbb{E}(X_k)\right).$$

Puisque X_1 et X_2 sont indépendantes, elles sont décorrélées et donc $\mathbb{V}\mathrm{ar}(X)$ est diagonale par bloc :

$$\mathbb{V}\mathrm{ar}(X) = \begin{pmatrix} \mathbb{V}\mathrm{ar}(X_1) & 0 \\ 0 & \mathbb{V}\mathrm{ar}(X_2) \end{pmatrix}$$

et on vérifie alors que

$$\sum_{k=1,2} t_k^T\mathbb{V}\mathrm{ar}(X_k)t_k = t^T\mathbb{V}\mathrm{ar}(X)t \quad \text{et} \quad \sum_{k=1,2} t_k^T\mathbb{E}(X_k) = t^T\mathbb{E}(X)$$

ce qui donne le résultat voulu.

Supposons maintenant que (X_1, X_2) est un vecteur gaussien et que X_1 et X_2 sont décorrélées : alors $\mathbb{V}\mathrm{ar}(X)$ est diagonale par bloc et coïncide donc, comme on vient de le voir, avec la variance d'un couple de vecteurs gaussiens indépendants. Puisque, à espérance fixée, la variance caractérise la loi, cela implique bien que X_1 et X_2 sont indépendantes. ■

4.2 Théorème central limite multi-dimensionnel

On généralise maintenant le théorème central limite 3.4.7 au cas multi-dimensionnel et sous une forme souvent utile en pratique.

Théorème 4.2.1 (Théorème central limite)**.** *Soit $(X_n, n \geq 1)$ à valeurs dans $\mathbb{R}^N$, i.i.d. avec $\mathbb{E}(X_1) = m$ et X_1 de carré intégrable et de matrice de covariance $\mathbb{V}\mathrm{ar}(X_1)$. Alors*

$$\frac{1}{n^{1/2}}\sum_{k=1}^{n}(X_k - m) \xrightarrow{\mathrm{L}} X$$

où X est un vecteur gaussien de moyenne nulle et de matrice de covariance $\mathbb{V}\mathrm{ar}(X_1)$.

Une ébauche de preuve de ce résultat suit l'ébauche de preuve du Théorème 3.4.7 où l'on remplace les produits par des produits scalaires.

4.3 Interprétation géométrique

On adopte dans cette partie un point de vue géométrique qui permet d'établir des propriétés fines sur les vecteurs gaussiens.

4.3.1 Décomposition spectrale de $\mathbb{V}\mathrm{ar}(X)$

On définit $I = \mathrm{Im}(\mathbb{V}\mathrm{ar}(X))$ et $r = \dim(I)$ l'image et le rang de $\mathbb{V}\mathrm{ar}(X)$, respectivement. On rappelle que $\mathbb{V}\mathrm{ar}(X)$ est symétrique positive, cf. Proposition 1.4.7 : en particulier, on peut la diagonaliser dans une base orthonormale $(V_1, \dots, V_n)$, toutes ses valeurs propres $\lambda_1, \dots, \lambda_n$ sont ≥ 0 et le nombre de valeurs propres non nulles est égale à son rang r. On écrira par la suite $\mathbb{V}\mathrm{ar}(X) = V\Delta V^T$ avec $V = (V_1 \ \cdots \ V_n)$ unitaire, i.e., $V^{-1} = V^T$, et Δ diagonale avec $\Delta_{ii} = \lambda_i$. Quitte à permuter les vecteurs de la base orthonormale, on supposera que $\lambda_i > 0$ si $i \in \{1, \dots, r\}$ et $\lambda_i = 0$ si $i \in \{r+1, \dots, n\}$. On notera enfin que par définition, I est l'espace engendré par $(V_1, \dots, V_r)$ et $\mathrm{Ker}(\mathbb{V}\mathrm{ar}(X))$, le noyau de $\mathbb{V}\mathrm{ar}(X)$, est l'espace engendré par $(V_{r+1}, \dots, V_n)$.

4.3.2 Projection sur l'image de $\mathbb{V}\mathrm{ar}(X)$

Puisque V est unitaire, i.e., $V^{-1} = V^T$, la décomposition $V\Delta V^T = \mathbb{V}\mathrm{ar}(X)$ donne $V^T\mathbb{V}\mathrm{ar}(X)V = \Delta$ puis, par définition de la matrice de variance, $\mathbb{V}\mathrm{ar}(V^T X) = \Delta$. Ce résultat élémentaire a la conséquence très importante suivante.

Théorème 4.3.1. *Si X est un vecteur gaussien, alors $X - \mathbb{E}(X)$ appartient presque sûrement à l'image de $\mathbb{V}\mathrm{ar}(X)$.*

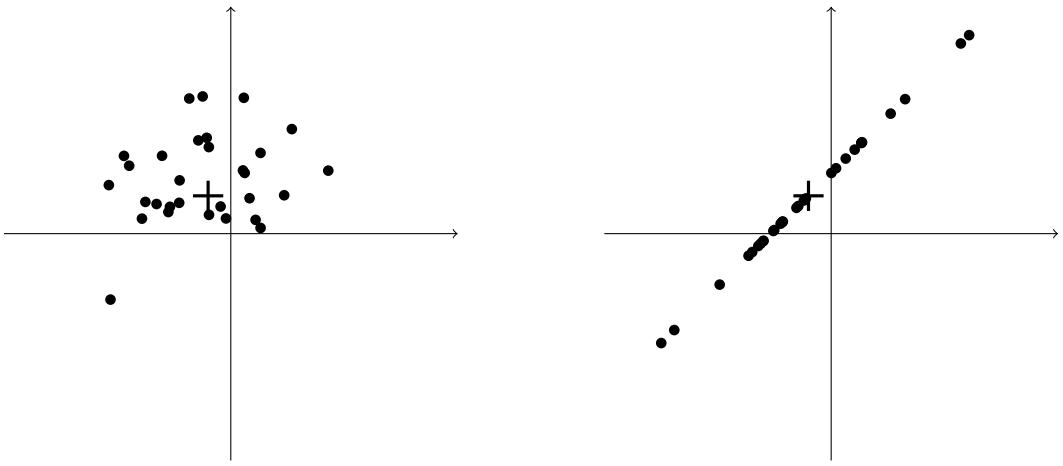
Démonstration. Partant de $\mathbb{V}\mathrm{ar}(V^T X) = \Delta$ et multipliant par e_k à droite et par e_k^T à gauche, il vient $\mathbb{V}\mathrm{ar}((Ve_k)^T X) = \lambda_k$. Puisque $Ve_k = V_k$ par définition et que $\lambda_k = 0$ pour $k = r+1, \dots, n$, on a donc prouvé que

$$\mathbb{V}\mathrm{ar}(V_k^T X) = 0, \ \ k = r+1, \dots, n.$$

La Proposition 1.3.10 implique donc que $V_k^T X = V_k^T \mathbb{E}(X)$ (presque sûrement) puis que $V_k^T(X - \mathbb{E}(X)) = 0$. Puisque cette relation est valable pour tout $k = r+1, \dots, n$, cela entraîne que $X - \mathbb{E}(X)$ appartient à l'orthogonal de l'espace engendré par $(V_{r+1}, \dots, V_n)$ qui n'est autre, par construction, que I. ■

Une conséquence de ce résultat est que si $\mathbb{V}\mathrm{ar}(X)$ n'est pas de rang plein, i.e., $r < n$, alors $X - \mathbb{E}(X)$ vit dans un espace vectoriel de dimension $< n$: ce résultat est illustré en dimension $n = 2$ sur la Figure 4.2.

En outre, il existe donc $Y \in \mathbb{R}^n$ tel que $X = \mathbb{V}\mathrm{ar}(X)Y + \mathbb{E}(X)$. Sauf lorsque $\mathbb{V}\mathrm{ar}(X)$ est de rang plein, Y n'est pas unique et peut être modifié par un vecteur du noyau de $\mathbb{V}\mathrm{ar}(X)$. Intuitivement, il y a r degrés de liberté et le résultat suivant montre qu'on peut effectivement se ramener à un vecteur de dimension r. Le résultat suivant est une généralisation de la Proposition 2.5.5 qui permet de se ramener au cas standard.



(a) Cas $\mathbb{V}\mathrm{ar}(X) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ (b) Cas $\mathbb{V}\mathrm{ar}(X) = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$

FIGURE 4.2 – Les deux figures présentent un échantillon de 30 vecteurs gaussiens (représentés par des points) en dimension $n = 2$: chaque point représente un vecteur gaussien, et dans chaque cas les 30 vecteurs gaussiens sont i.i.d. de moyenne $\left(\begin{smallmatrix} -0{,}3 \\ 0{,}5 \end{smallmatrix}\right)$. Les deux figures différent par la matrice de variance choisie : pour la Figure 4.2a on a pris $\mathbb{V}\mathrm{ar}(X) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ et $\mathbb{V}\mathrm{ar}(X) = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ pour la Figure 4.2b. Dans chaque cas, les points vivent dans un espace dont la dimension est égale au rang de $\mathbb{V}\mathrm{ar}(X)$: ainsi, lorsque $\mathbb{V}\mathrm{ar}(X)$ n'est pas de rang plein, les vecteurs gaussiens vivent dans un espace de dimension plus petite (ici, une droite).

Théorème 4.3.2. *Soit X un vecteur gaussien. Il existe une décomposition de la forme $X = WY + \mathbb{E}(X)$ où :*

- *$W \in \mathbb{R}^{n\times r}$ a la même image que $\mathbb{V}\mathrm{ar}(X)$: $\mathrm{Im}(W) = \mathrm{Im}(\mathbb{V}\mathrm{ar}(X))$;*
- *$Y \in \mathbb{R}^d$ est un vecteur gaussien standard.*

Cette décomposition est donnée de manière explicite par

$$W = \tilde{V}\tilde{\Delta}^{1/2} \in \mathbb{R}^{n\times r} \quad \text{et} \quad Y = \tilde{\Delta}^{-1/2}\tilde{V}^T(X - \mathbb{E}(X)) \in \mathbb{R}^d$$

où :

- *$\tilde{\Delta} \in \mathbb{R}^{r\times r}$ est la matrice diagonale avec $\tilde{\Delta}_{ii} = \lambda_i$ pour $i \in \{1, \dots, r\}$;*
- *$\tilde{V} = \begin{pmatrix} V_1 & \cdots & V_r \end{pmatrix}$ est la matrice de taille $n \times r$ obtenue en gardant les r premières colonnes de V.*

Démonstration. On commence par vérifier que Y est un vecteur gaussien standard. Tout d'abord, Y est bien un vecteur gaussien comme image d'un vecteur gaussien par une transformation affine. En outre, il est centré puisque $\mathbb{E}(Y) = \tilde{\Delta}^{-1/2}\tilde{V}^T\mathbb{E}(X - \mathbb{E}(X)) = 0$, et on calcule sa variance :

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(Y) &= \mathbb{V}\mathrm{ar}\left(\tilde{\Delta}^{-1/2}\tilde{V}^T(X - \mathbb{E}(X))\right) && \text{(définition de } Y\text{)} \\
&= \mathbb{V}\mathrm{ar}\left(\tilde{\Delta}^{-1/2}\tilde{V}^T X\right) && \text{(invariance de la variance par translation)} \\
&= \tilde{\Delta}^{-1/2}\tilde{V}^T\mathbb{V}\mathrm{ar}\,(X)\,\tilde{V}\tilde{\Delta}^{-1/2} && \text{(déf. de la variance et linéarité de l'espérance).}
\end{aligned}$$

Par définition, on a la décomposition par bloc

$$\Delta = \begin{pmatrix} \tilde{\Delta} & 0 \\ 0 & 0 \end{pmatrix} \quad \text{et} \quad V = \begin{pmatrix} \tilde{V} & \hat{V} \end{pmatrix} \quad \text{avec} \quad \hat{V} = \begin{pmatrix} V_{r+1} & \cdots & V_n \end{pmatrix}$$

et donc

$$\mathbb{V}\mathrm{ar}(X) = V\Delta V^T = \begin{pmatrix} \tilde{V} & \hat{V} \end{pmatrix} \begin{pmatrix} \tilde{\Delta} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \tilde{V}^T \\ \hat{V}^T \end{pmatrix} = \tilde{V}\tilde{\Delta}\tilde{V}^T$$

ce qui donne

$$\mathbb{V}\mathrm{ar}(Y) = \tilde{\Delta}^{-1/2}\tilde{V}^T\tilde{V}\tilde{\Delta}\tilde{V}^T\tilde{V}\tilde{\Delta}^{-1/2} = I_d.$$

On a donc bien montré que Y est un vecteur gaussien standard, et il reste à montrer que $X = WY + \mathbb{E}(X)$ avec $\mathrm{Im}(W) = I$. On a bien $\mathrm{Im}(W) = I$: en effet, $\tilde{\Delta}$ étant inversible on a $\mathrm{Im}(W) = \mathrm{Im}(\tilde{V})$ qui est l'espace engendré par ses colonnes $(V_1, \ldots, V_r)$, qui est bien I. Enfin, on calcule

$$WY = \tilde{V}\tilde{\Delta}^{1/2}\tilde{\Delta}^{-1/2}\tilde{V}^T(X - \mathbb{E}(X)) = \tilde{V}\tilde{V}^T(X - \mathbb{E}(X)).$$

Pour $x \in \mathrm{Im}(\mathbb{V}\mathrm{ar}(X))$, on a $\tilde{V}\tilde{V}^T x = x$ ce qui donne le résultat puisque $X \in I$ presque sûrement par le Théorème 4.3.1. ■

4.3.3 Absolue continuité

Corollaire 4.3.3. *Si $\mathbb{V}\mathrm{ar}(X)$ n'est pas inversible, alors X n'est pas absolument continue.*

Démonstration. Si $X \in \mathbb{R}^n$ est absolument continue, alors d'après la Remarque 2.5.4 on a $\mathbb{P}(X - \mathbb{E}(X) \in E) = 0$ pour tout espace vectoriel de dimension $\dim(E) < n$. Puisque $\mathbb{P}(X - \mathbb{E}(X) \in \mathrm{Im}(\mathbb{V}\mathrm{ar}(X))) = 1$ d'après le Théorème 4.3.1, il s'ensuit que $\dim(\mathrm{Im}(\mathbb{V}\mathrm{ar}(X))) = n$ lorsque X est absolument continue. ■

Le Théorème 4.3.2 implique que la réciproque est vraie, et on peut alors calculer la densité de X.

Théorème 4.3.4. *Soit X un vecteur gaussien. Si $\mathbb{V}\mathrm{ar}(X)$ est inversible, alors X est absolument continue et sa densité est donnée par*

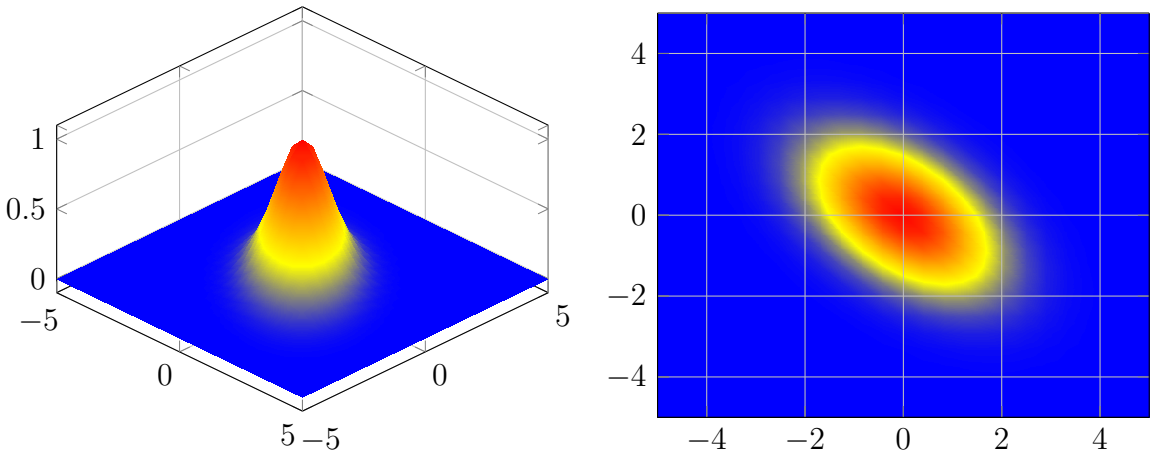
$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n \det(\mathbb{V}\mathrm{ar}(X))}} \exp\left(-\frac{1}{2}(x - \mathbb{E}(X))^T \mathbb{V}\mathrm{ar}(X)^{-1}(x - \mathbb{E}(X))\right), \quad x \in \mathbb{R}^n.$$

Démonstration. Lorsque $r = n$, le Théorème 4.3.2 implique que $Y = \Delta^{-1/2}V^T(X - \mathbb{E}(X))$ est un vecteur gaussien standard en dimension n, et que $X = V\Delta^{1/2}Y + \mathbb{E}(X)$. Puisque Y est absolument continue et que l'application $x \mapsto \Delta^{-1/2}V^T x + \mathbb{E}(X)$ est infiniment différentiable et que son jacobien vaut $\Delta^{-1/2}V^T$, le Théorème 2.5.1 montre que X est absolument continue et que

$$f_X(x) = f_Y\left(\Delta^{-1/2}V(x - \mathbb{E}(X))\right) \left|\det\left(\Delta^{-1/2}V^T\right)\right|.$$

D'un côté, la formule (4.1) implique que

$$\begin{aligned} f_Y\left(\Delta^{-1/2}V(x - \mathbb{E}(X))\right) &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}(x - \mathbb{E}(X))^T V\Delta^{-1/2}\Delta^{-1/2}V^T(x - \mathbb{E}(X))\right) \\ &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}(x - \mathbb{E}(X))^T \mathbb{V}\mathrm{ar}(X)^{-1}(x - \mathbb{E}(X))\right) \end{aligned}$$



(a) Vue de profil.

(b) Vue de haut.

FIGURE 4.3 – Densité de la loi normale sur $\mathbb{R}^2$ de moyenne $\mathbb{E}(X) = \left(\begin{smallmatrix}1\\2\end{smallmatrix}\right)$ et de variance $\mathbb{V}\mathrm{ar}(X) = \left(\begin{smallmatrix}1 & \\ 2 & 1\end{smallmatrix}\right)$ vue de deux angles différents. Pour revenir à la loi normale standard, il faut opérer une rotation et dilater les axes, ce qui correspond exactement au Théorème 4.3.2.

et d'un autre côté, puisque V est unitaire on

$$\left|\det\left(\Delta^{-1/2}V^T\right)\right| = \left|\det\left(\Delta\right)\right|^{-1/2} = \det(\mathbb{V}\mathrm{ar}(X))^{-1/2}.$$

Cela prouve le résultat. ■

Dans le cas standard, les courbes d'iso-densité sont des cercles : dans le cas général, ce sont donc des ellipses de la forme $\{x : (x - \mathbb{E}(X))^T \mathbb{V}\mathrm{ar}(X)^{-1}(x - \mathbb{E}(X)) = c\}$.

4.4 Définition standard

La caractéristation suivante d'un vecteur gaussien comme unique vecteur dont toute combinaison linéaire des coordonnées suit une loi normale sur $\mathbb{R}$ est très importante d'un point de vue conceptuel et est par ailleurs généralement prise comme définition d'un vecteur gaussien.

Théorème 4.4.1. *$X \in \mathbb{R}^n$ est un vecteur gaussien si et seulement si pour tout $t \in \mathbb{R}^n$, $t^T X$ suit une loi normale sur $\mathbb{R}$.*

Démonstration. Soit X un vecteur gaussien : on peut donc écrire $X = MY + \mu$ comme dans la Définition 4.1.2, si bien que $t^T X = t^T MY + t^T \mu$. Pour montrer que $t^T X$ suit une loi normale, il suffit de montrer que sa fonction caractéristique est celle d'une loi normale donnée par (2.5.6). En utilisant la proposition précédente, on calcule

$$\varphi_{t^T X}(a) = \mathbb{E}(e^{iat^T X}) = \varphi_X(at^T) = \exp\left(-\frac{1}{2}a^2 t^T \mathbb{V}\mathrm{ar}(X)t + iat^T \mathbb{E}(X)\right)$$

et donc $\varphi_{t^T X}$ est bien de la forme attendue. Pour montrer la réciproque, on suppose que pour tout $t \in \mathbb{R}^n$, $t^T X$ suit une loi normale sur $\mathbb{R}$: en particulier, la Proposition (2.5.6) montre que

$$\mathbb{E}(e^{it^T X}) = \varphi_{t^T X}(1) = \exp\left(-\frac{1}{2}\mathbb{V}\mathrm{ar}(t^T X) + i\mathbb{E}(t^T X)\right).$$

Par linéarité de l'espérance et définition de la variance, $\mathbb{E}(t^T X) = t^T \mathbb{E}(X)$ et

$$\mathbb{V}\mathrm{ar}(t^T X) = \mathbb{E}(t^T (X - \mathbb{E}(X))(X - \mathbb{E}(X))^T t) = t^T \mathbb{E}((X - \mathbb{E}(X))(X - \mathbb{E}(X))^T) t = t^T \mathbb{V}\mathrm{ar}(X) t$$

et donc

$$\mathbb{E}(e^{it^T X}) = \exp\left(-\frac{1}{2} t^T \mathbb{V}\mathrm{ar}(X) t + i t^T \mathbb{E}(X)\right).$$

Puisque $\mathbb{E}(e^{it^T X}) = \varphi_X(t)$ et que le membre de droite est la fonction caractéristique du vecteur gaussien de moyenne $\mathbb{E}(X)$ et de matrice de covariance $\mathbb{V}\mathrm{ar}(X)$, cela montre bien que X est un vecteur gaussien. ■

4.5 Espérance conditionnelle

De manière générale, $\mathbb{E}(X_2 \mid X_1) = h(X_1)$ avec h une fonction mesurable. Dans le cas où (X_1, X_2) est un vecteur gaussien avec $\mathbb{V}\mathrm{ar}(X_1)$ inversible, le résultat suivant montre que h est tout simplement une fonction affine.

Théorème 4.5.1. *Soit $X = (X_1, X_2)$ un vecteur gaussien tel que X_1 est absolument continu. Alors*

$$\mathbb{E}(X_2 \mid X_1) = \mathbb{E}(X_2) + \mathbb{C}\mathrm{ov}(X_2, X_1)\mathbb{V}\mathrm{ar}(X_1)^{-1} (X_1 - \mathbb{E}(X_1)).$$

En particulier, $\mathbb{E}(X_2 \mid X_1)$ est un vecteur gaussien de matrice de moyenne $\mathbb{E}(X_2)$ et de matrice de covariance $\mathbb{C}\mathrm{ov}(X_2, X_1)\mathbb{V}\mathrm{ar}(X_1)^{-1}\mathbb{C}\mathrm{ov}(X_1, X_2)$.

Démonstration. Soit $Y_i = X_i - \mathbb{E}(X_i)$ pour $i = 1, 2$: afin de montrer le résultat, il suffit de montrer que $\mathbb{E}(Y_2 \mid Y_1) = \mathbb{C}\mathrm{ov}(Y_2, Y_1)\mathbb{V}\mathrm{ar}(Y_1)^{-1} Y_1$, ce qui nous permet donc de nous ramener au cas centré. Pour cela, l'idée est de chercher une matrice M telle que $Y_2 - MY_1$ et Y_1 soient indépendants. Puisque $(Y_2 - MY_1, Y_1)$ est un vecteur gaussien (comme transformation linéaire du vecteur gaussien (Y_1, Y_2)), il suffit d'annuler la covariance entre $Y_2 - MY_1$ et Y_1 pour qu'ils soient indépendants : il s'agit donc de résoudre

$$0 = \mathbb{C}\mathrm{ov}(Y_2 - MY_1, Y_1) = \mathbb{E}\left[(Y_2 - MY_1) Y_1^T\right] = \mathbb{C}\mathrm{ov}(Y_2, Y_1) - M\mathbb{V}\mathrm{ar}(Y_1).$$

Puisque $\mathbb{V}\mathrm{ar}(Y_1)$ est inversible, il suffit de prendre $M = \mathbb{C}\mathrm{ov}(Y_2, Y_1)\mathbb{V}\mathrm{ar}(Y_1)^{-1}$. On définit alors $Z = Y_2 - MY_1$: alors $Y_2 = MY_1 + Z$ et donc

$$\mathbb{E}(Y_2 \mid Y_1) = \mathbb{E}(MY_1 \mid Y_1) + \mathbb{E}(Z \mid Y_1).$$

On a $\mathbb{E}(MY_1 \mid Y_1) = MY_1$ et $\mathbb{E}(Z \mid Y_1) = \mathbb{E}(Z)$ puisque Z et Y_1 sont indépendants, cf. Proposition 1.7.1. Puisque $\mathbb{E}(Z) = 0$ on obtient bien le résultat. ■

Ce résultat a de nombreuses conséquences. On notera par exemple la formule élégante suivante,.

Corollaire 4.5.2. *Si (X_1, X_2, X_3) est un vecteur gaussien, que $\mathbb{E}(X_1) = 0$ et que X_2 et X_3 sont indépendants et absolument continus, alors*

$$\mathbb{E}(X_1 \mid X_2, X_3) = \mathbb{E}(X_1 \mid X_2) + \mathbb{E}(X_1 \mid X_3).$$

Démonstration. Une première preuve consiste à utiliser le Théorème 4.5.1 puis effectuer des calculs matriciels. On propose ici une preuve géométrique.

On suppose sans perte de généralité que $\mathbb{E}(X_2) = \mathbb{E}(X_3) = 0$ et on définit $\mathrm{Vect}(Z) = \{AZ : A \in \mathbb{R}^{n_1 \times d}\}$ pour n'importe quelle variable aléatoire Z de dimension d, et où n_i est la dimension de X_i. De manière générale, $\mathbb{E}(X_1 \mid X_2, X_3)$ est la projection orthogonale de X_1 sur l'espace engendré par (X_2, X_3), i.e., sur l'espace $\{h(X_2, X_3) : h : \mathbb{R}^{n_2+n_3} \to \mathbb{R}^{n_1} \text{ mesurable}\}$, cf. Section 1.7.5. Dans le cas de vecteurs gaussiens, le Théorème 4.5.1 montre qu'en fait, on peut se restreindre à la projection sur l'espace engendré par les transformations linéaires de (X_2, X_3), i.e., on projette X_1 sur $\mathrm{Vect}(X_2, X_3)$. L'indépendance et le fait que X_2 et X_3 sont centrées impliquent alors que $\mathrm{Vect}(X_2, X_3) = \mathrm{Vect}(X_2) \oplus \mathrm{Vect}(X_3)$ ce qui donne le résultat : de manière générale, la projection orthogonale d'un élément sur l'union directe de deux espaces orthogonaux est égale à la somme des projections orthonogales sur ces deux sous-espaces. ■

4.6 Indépendance des moyenne et variance empiriques, loi du χ^2 et loi de Student

On conclut ce chapitre par un résultat qui joue un rôle prépondérant en statistique. Ce résultat fait intervenir la loi du χ^2 et la loi de Student qui sont deux mesures de probabilité liées aux vecteurs gaussiens.

Définition 4.6.1. La loi du χ^2 à $r \in \mathbb{N}^*$ degrés de liberté est la loi de la variable aléatoire $X^T X$ où X est un vecteur gaussien standard en dimension n.

Derrière tout vecteur gaussien se cache un vecteur gaussien standard, cf. Théorème 4.3.2. Ainsi, la loi du χ^2 apparaît naturellement comme suit.

Proposition 4.6.1. *Soit X un vecteur gaussien en dimension n tel que $\mathbb{V}\mathrm{ar}(X)$ est inversible. Alors $(X - \mathbb{E}(X))^T \mathbb{V}\mathrm{ar}(X)^{-1}(X - \mathbb{E}(X))$ suit la loi du χ^2 à $r \in \mathbb{N}^*$ degrés de liberté.*

Démonstration. En utilisant les notations du Théorème 4.3.2, on voit que

$$(X - \mathbb{E}(X))^T \mathbb{V}\mathrm{ar}(X)^{-1}(X - \mathbb{E}(X)) = Y^T \Delta^{1/2} V^{1/2} V^T \Delta^{-1} V \mathbb{V}\mathrm{ar}(X)^{1/2} \Delta^{1/2} Y = Y^T Y$$

ce qui donne le résultat. ■

Définition 4.6.2. La loi de Student à $r \in \mathbb{N}^*$ degrés de liberté est la loi de la variable aléatoire

$$\frac{Z}{\sqrt{V_r/r}}$$

où Z est une variable aléatoire normale standard, V_r suit une loi du χ^2 à r degrés de liberté et Z et V_r sont indépendantes.

Il découle directement de ces définitions que si V_r suit la loi du χ^2 à r degrés de libertés, alors V_r/r converge en loi vers un lorsque $r \to \infty$. En particulier, si S_r suit la loi de Student à r degrés de liberté alors S_r converge en loi lorsque $r \to \infty$ vers la loi normale standard (cf. Propositions 3.4.3 et 3.4.4).

Théorème 4.6.2. *Si $X_1, \ldots, X_n$ sont n variables aléatoires gaussiennes réelles i.i.d. de moyenne $m \in \mathbb{R}$ et de variance $\sigma^2 \in \mathbb{R}_+$, alors les variables aléatoires*

$$\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k \quad \text{et} \quad S_{n-1}^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \bar{X}_n\right)^2$$

sont indépendantes. Par ailleurs,

- *$\bar{X}_n$ suit une loi normale sur $\mathbb{R}$, de moyenne m et de variance σ^2/n ;*
- *$(n-1)S_{n-1}^2/\sigma^2$ suit la loi du χ^2 à $n-1$ degrés de liberté ;*

et donc

$$n^{1/2}\frac{\left(\bar{X}_n - m\right)}{S_{n-1}}$$

suit la loi de Student à $n-1$ degrés de liberté.

4.7 Fiche de synthèse

Loi normale $\mathcal{N}(m, \sigma^2)$ sur $\mathbb{R}$

- de densité $f : x \in \mathbb{R} \mapsto \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$;
- de fonction caractéristique $\varphi : t \in \mathbb{R} \mapsto \exp\left(itm - \frac{\sigma^2 t^2}{2}\right)$.

Définition : $X = (X_1, \ldots, X_n)$ est un **vecteur gaussien standard** en dimension n si les X_i sont i.i.d. et suivent des lois normales standard $\mathcal{N}(0, 1)$.

Définition : $X \in \mathbb{R}^n$ est un **vecteur gaussien** s'il existe $Y \in \mathbb{R}^m$ vecteur gaussien standard, $M \in \mathbb{R}^{n\times m}$ et $\mu \in \mathbb{R}^n$ tels que $X = MY + \mu$, i.e., si X est l'image par une transformation affine d'un vecteur gaussien standard.

Remarque : En particulier, chaque coordonnée X_i suit une loi normale : par contre la réciproque n'est pas vraie en général, sauf sous une hypothèse supplémentaire d'indépendance.

Propriété : Si les X_i sont indépendantes de lois normales, alors X est un vecteur gaussien.

Propriété : X est un vecteur gaussien si et seulement si pour tout $t \in \mathbb{R}^n$ $t^T X$ suit une loi normale (sur $\mathbb{R}$).

Densité et fonction caractéristique d'un vecteur gaussien :

- Si $\mathbb{V}\mathrm{ar}(X)$ n'est pas inversible, alors X n'est pas absolument continu.
- Si $\mathbb{V}\mathrm{ar}(X)$ est inversible, alors X est absolument continu, de densité f_X où

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^d \mathrm{dét}(\mathbb{V}\mathrm{ar}(X))}} \exp\left(-\frac{1}{2}(x - \mathbb{E}(X))^T \mathbb{V}\mathrm{ar}(X)^{-1}(x - \mathbb{E}(X))\right).$$

où $\mathbb{E}(X)$ est le vecteur $(\mathbb{E}(X_i))_i$ et $\mathbb{V}\mathrm{ar}(X)$ la matrice de covariance $(\mathbb{C}\mathrm{ov}(X_i, X_j))_{i,j}$.

Propriété : Les X_i sont indépendantes si et seulement si $\mathbb{V}\mathrm{ar}(X)$ est une matrice diagonale.

Conditionnement :
Si $X = (X_1, X_2)$ est un vecteur gaussien tel que X_1 est absolument continu, alors

$$\mathbb{E}(X_2 \mid X_1) = \mathbb{E}(X_2) + \mathbb{C}\mathrm{ov}(X_2, X_1)\mathbb{V}\mathrm{ar}(X_1)^{-1}(X_1 - \mathbb{E}(X_1)).$$

En particulier, $\mathbb{E}(X_2 \mid X_1)$ est un vecteur gaussien de matrice de moyenne $\mathbb{E}(X_2)$ et de matrice de covariance $\mathbb{C}\mathrm{ov}(X_2, X_1)\mathbb{V}\mathrm{ar}(X_1)^{-1}\mathbb{C}\mathrm{ov}(X_1, X_2)$.

Corollaire : Si (X_1, X_2, X_3) est un vecteur gaussien tel que $\mathbb{E}(X_1) = 0$ et que X_2 et X_3 sont indépendants et absolument continus, alors

$$\mathbb{E}(X_1 \mid X_2, X_3) = \mathbb{E}(X_1 \mid X_2) + \mathbb{E}(X_1 \mid X_3).$$

4.8 Exercices

Les exercices précédés d'une flèche ↪ sont des exercices d'application directs du cours.

↪ **Exercice 4.1**

1. Soit $\mathbf{X} = (X_1, X_2, X_3)$ le vecteur gaussien centré de matrice de covariance

$$\mathbb{V}\mathrm{ar}(\mathbf{X}) = \begin{pmatrix} 3 & 6 & 2 \\ 6 & 14 & 3 \\ 2 & 3 & 6 \end{pmatrix}$$

Calculez $\mathbb{E}(X_3 \mid X_1, X_2)$ à l'aide du théorème 4.5.1 du cours.

Dans le reste de l'exercice on propose une autre méthode pour calculer cette espérance conditionnelle. On considère $\mathbf{X}_n = (X_1, \dots, X_n)$ un vecteur gaussien centré.

2. Montrez que pour tout $i = 1, \dots, n-1$, on a

$$\mathbb{E}(X_i X_n) = \mathbb{E}\big(X_i \mathbb{E}(X_n \mid X_1, \dots, X_{n-1})\big).$$

3. En déduire un système linéaire de $n-1$ équations dont la solution permet d'exprimer $\mathbb{E}(X_n \mid X_1, \dots, X_{n-1})$ comme une combinaison linéaire de $X_1, \dots, X_{n-1}$.

4. Retrouvez le résultat de la première question par cette méthode.

↪ **Exercice 4.2** *(Coordonnées polaires)*

Soient X, Y deux variables aléatoires réelles indépendantes de densité de probabilité f_X et f_Y. On considère le changement de variables en coordonnées polaires $X = R\cos(\Theta)$ et $Y = R\sin(\Theta)$ avec $R \in [0, \infty)$ et $\Theta \in [0, 2\pi]$.

1. Calculer en fonction des densités f_X et f_Y la loi de (R, Θ) et en déduire la loi de R.

2. En déduire que si X et Y suivent des lois gaussiennes centrées réduites, alors R et Θ sont indépendantes.

↪ **Exercice 4.3** *(Loi du χ^2)*

Soit X un vecteur gaussien centré en dimension n avec $\mathbb{V}\mathrm{ar}(X)$ inversible : on prouve la Proposition 4.6.1 d'une manière différente.

1. Montrez que la transformée de Laplace L_r de la loi du χ^2 à r degrés de liberté est donnée par

$$L_r(t) = \frac{1}{(1+2t)^{r/2}}.$$

2. Utilisez le théorème de transfert pour calculer la transformée de Laplace $L_{X\mathbb{V}\mathrm{ar}(X)^{-1}X^T}$ de $X\mathbb{V}\mathrm{ar}(X)^{-1}X^T$.

3. Déduisez-en que $X\mathbb{V}\mathrm{ar}(X)^{-1}X$ suit une loi du χ^2.

Exercice 4.4

Soit Θ_i i.i.d. dans $[0, 2\pi]$, $X_i = \cos\Theta_i$ et $Y_i = \sin\Theta_i$.

1. Montrez que $\mathbb{E}(X_i) = \mathbb{E}(Y_i) = \mathbb{E}(X_i Y_i)$ et déduisez-en que $\mathbb{C}\mathrm{ov}(X_i, Y_i) = 0$.

2. Montrez que les variables X_i et Y_i ne sont pas indépendantes.

3. Soit $\bar{X}_n = n^{-1/2}(X_1 + \cdots + X_n)$ et $\bar{Y}_n = n^{-1/2}(Y_1 + \cdots + Y_n)$. Calculez $\mathbb{C}\mathrm{ov}(\bar{X}_n, \bar{Y}_n)$. Les variables aléatoires $\bar{X}_n$ et $\bar{Y}_n$ sont-elles indépendantes ?

4. Montrez que $(\bar{X}_n, \bar{Y}_n) \xrightarrow{\mathrm{L}} (X_\infty, Y_\infty)$ avec (X_∞, Y_∞) que l'on identifiera. Justifiez que X_∞ et Y_∞ sont indépendantes.

Exercice 4.5

Le but de ce problème est de prouver le Théorème 4.6.2. Sans perte de généralité on supposera que $m = 0$ et $\sigma^2 = 1$.

1. Calculez $\mathbb{C}\text{ov}(X_1 + X_2, X_1 - X_2)$ et déduisez-en le résultat pour $n = 2$.

On montre maintenant le résultat dans le cas général. Pour cela, $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^n$, $A_1 = (1/n)\mathbf{1}\mathbf{1}^T$ la matrice de taille $n \times n$ dont toutes les entrées valent $1/n$ et $A_2 = I - A_1$ avec I la matrice identité de taille $n \times n$.

2. Montrez que $n\bar{X}_n^2 = X^T A_1 X$ et que $(n-1)S_{n-1}^2 = X^T A_2 X$.

3. Justifiez que l'on puisse écrire $A_1 = U^T \Delta U$ et $A_2 = U^T(I - \Delta)U$ avec U unitaire et $\Delta_{11} = 1$ et $\Delta_{ij} = 0$ sinon.

4. Montrez que UX est un vecteur gaussien standard et concluez.

Deuxième partie

Statistique

Chapitre 5

Estimation paramétrique

5.1 Introduction

Afin de motiver le concept d'estimation paramétrique, nous commencerons par discuter un exemple illustratif simple.

5.1.1 Description informelle

Imaginons la situation concrète suivante. J'ai à ma disposition une pièce de monnaie dont j'aimerais identifier le biais (= probabilité d'obtenir pile) : comment faire ? Après des discussions endiablées lors du dernier gala de l'école, mes amis et moi avons convenu de la procédure suivante : je vais jeter la pièce 103 fois puis je vais enregistrer le nombre de succès, i.e., le nombre de fois où j'aurai obtenu pile. Si N est ce nombre, alors je considérerai $N/103$ comme estimation du biais de ma pièce.

5.1.2 Formalisation

Nous passons maintenant à la formalisation de la description ci-dessus. Comme on l'a discuté en détail au Chapitre 1, l'expérience aléatoire consistant à lancer une pièce de monnaie peut être décrite par l'espace de probabilités $(\Omega, \mathcal{F}, \mathbb{P})$ où $\Omega = \{0, 1\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ et $\mathbb{P}$ décrit les caractéristiques statistiques de la pièce : la probabilité sous $\mathbb{P}$ de l'évènement élémentaire $\{1\}$, que l'on notera $\theta = \mathbb{P}(\{1\})$, représente la probabilité d'obtenir pile lors de l'expérience aléatoire. Dans l'exemple précédent, **on cherche donc à estimer le paramètre inconnu** θ.

Formalisons maintenant la procédure d'estimation proposée ci-dessus : jeter n fois la pièce ($n = 103$ dans l'exemple ci-dessus) correspond à considérer n variables aléatoires $X_1, \ldots, X_n$ de même loi $\mathbb{P}$ et indépendantes, et le nombre de fois où l'on a obtenu pile, noté N_n, est donc donné par

$$N_n = \sum_{k=1}^{n} \mathbb{1}\{X_k = 1\}.$$

Puisque les variables aléatoires $(\mathbb{1}\{X_k = i\}, k \in \mathbb{N}^*)$ sont i.i.d., intégrables et de moyenne θ, la loi des grands nombres assure $N_n/n \xrightarrow{\text{p.s.}} \theta$ et justifie donc d'utiliser $N_{103}/103$ comme estimation de θ. Quant à la question du choix de $n = 103$ (en pratique, on ne peut pas jeter la pièce une infinité de fois) c'est une question que l'on considérera dans la section de ce

chapitre dédiée aux intervalles de confiance.

Prenons maintenant un peu de hauteur et définissions $\Theta = [0, 1]$. Ainsi, chaque $\theta \in \Theta$ correspond à une mesure de probabilité $\mathbb{P}_\theta$ sur Ω qui n'est rien d'autre que la mesure de Bernoulli de paramètre $\theta = \mathbb{P}_\theta(\{1\})$. Dans l'exemple considéré, on connaît parfaitement l'expérience aléatoire mais pas ses propriétés statistiques : on sait qu'on jette une pièce de monnaie et on connaît donc l'espace mesurable $(\Omega, \mathcal{F})$ mais on ne connaît pas la mesure de probabilité $\mathbb{P}$ qui décrit cette expérience. Par contre, on sait que $\mathbb{P}$ fait partie de la famille de mesures de probabilités $\mathbb{P}_\theta$ indexées par Θ, ce que l'on notera $\mathbb{P} \in \{\mathbb{P}_\theta : \theta \in \Theta\}$, et on cherche en fait le "vrai" paramètre θ^* tel que $\mathbb{P} = \mathbb{P}_{\theta^*}$. On peut alors résumer le problème de l'estimation paramétrique de la manière suivante :

Parmi la famille $\{\mathbb{P}_\theta : \theta \in \Theta\}$, on cherche le paramètre $\theta \in \Theta$ qui rend le mieux compte des observations $X_1, \dots, X_n$.

Le paramètre choisi, noté $\hat{\theta}_n$, est appelé un estimateur de θ^* : $\hat{\theta}_n$ **est donc d'une fonction des observations** $X_1, \dots, X_n$ et l'on notera donc parfois $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ pour mettre l'accent sur le fait que

L'estimateur $\hat{\theta}_n$ est une variable aléatoire.

Dans l'exemple ci-dessus, on a donc fait le choix

$$\hat{\theta}_n^{(1)} = \frac{N_n}{n}$$

qui satisfait $\hat{\theta}_n^{(1)} \overset{\text{p.s.}}{\to} \theta^*$ par la loi forte des grands nombres. Néanmoins, si ce choix est naturel il n'est pas unique : par exemple, puisque l'on cherche le paramètre θ qui rend le mieux compte des observations $X_1, \dots, X_n$, un autre choix naturel est celui qui maximise la probabilité d'obtenir les valeurs observées :

$$\hat{\theta}_n^{(2)} \in \arg \max_{\theta \in \Theta} \mathcal{L}(\theta; X_1, \dots, X_n)$$

où $\mathcal{L}(\theta; x_1, \dots, x_n) = \mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n)$ pour $x_1, \dots, x_n \in \Omega$ est appelée vraisemblance : c'est la probabilité, si le biais de la pièce valait θ, d'obtenir $(x_1, \dots, x_n)$ comme résultat des n lancers. Le but de ce chapitre est de présenter une introduction à la théorie de ces estimateurs : on verra par exemple quel est le lien entre les deux estimateurs $\hat{\theta}_n^{(1)}$ et $\hat{\theta}_n^{(2)}$.

5.1.3 Un exemple en dimension $d > 1$

Dans l'exemple ci-dessus on a considéré un modèle paramétrique en dimension un, i.e., avec $\Theta \subset \mathbb{R}$. Par la suite, on considérera le cas général $\Theta \subset \mathbb{R}^d$. Un exemple naturel est l'estimation du paramètre d'un dé à 6 faces : dans ce cas, l'ensemble des paramètres Θ est donné par

$$\Theta = \left\{(p_1, \dots, p_5) \in [0, 1]^5 : p_1 + \cdots + p_5 \leq 1\right\}.$$

C'est un sous-ensemble de $\mathbb{R}^5$ qui correspond donc à un problème d'estimation en dimension $d = 5$.

5.2 Définitions et hypothèses

5.2.1 Modèle paramétrique et estimateurs

On considère un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$, un sous-ensemble mesurable $\Theta \subset \mathbb{R}^d$ pour un certain $d \in \mathbb{N}^*$ et $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ une famille de mesures de probabilités sur Ω : $\mathcal{P}$ est appelé **modèle paramétrique**. On supposera dans tout ce cours que le modèle est **identifiable**, c'est-à-dire que les mesures de probabilité de deux paramètres différents sont différentes :

$$\forall \theta, \theta' \in \Theta : \ \theta \neq \theta' \Rightarrow \mathbb{P}_\theta \neq \mathbb{P}_{\theta'}.$$

La mesure de probabilité $\mathbb{P}$ représente la "vraie" mesure de probabilité qui rend bien compte de l'expérience aléatoire à laquelle on s'intéresse, et on suppose que $\mathbb{P} \in \{\mathbb{P}_\theta : \theta \in \Theta\}$. Puisque le modèle est identifiable, il existe donc un unique paramètre $\theta^* \in \Theta$ tel que $\mathbb{P} = \mathbb{P}_{\theta^*}$ et le but est de trouver une bonne estimation de θ^* (et donc de $\mathbb{P}$). On s'intéressera parfois plus généralement à des fonctions de θ^* de la forme $g(\theta^*)$: par exemple, on cherchera parfois à estimer la moyenne de $\mathbb{P}$, donnée dans le cas discret par $\sum_x x\mathbb{P}(\{x\})$.

Exemple 5.2.1. Si on cherche à estimer le paramètre d'une loi exponentielle, on prendra $d = 1$, $\Theta =]0, \infty[$ et $\mathbb{P}_\theta$ la loi exponentielle de paramètre $\theta \in \Theta$, et donc de moyenne $1/\theta$.

Si on cherche une variable gaussienne en dimension un, on pourra prendre $d = 2$, $\Theta = \mathbb{R} \times \mathbb{R}_+$ et $\mathbb{P}_{m,\sigma}$ avec $(m, \sigma) \in \Theta$ la loi de la variable aléatoire gaussienne de moyenne m et de matrice de covariance σ. On vérifie aisément que ces deux modèles sont identifiables.

La donnée de base de notre problème est une suite $(X_n, n \in \mathbb{N}^*)$ de variables aléatoires i.i.d. de loi $\mathbb{P}$, que l'on appellera échantillon : le but est, à partir de cette suite, de construire un estimateur du paramètre θ^* qui nous intéresse. Evidemment, on ne connait pas $\mathbb{P}$ mais par contre, on connaît les lois $\mathbb{P}_\theta$ et l'on sera amenés à faire des calculs sous ces mesures de probabilités :

Faire des calculs sous $\mathbb{P}_\theta$ revient à faire des calculs en supposant que la loi commune aux X_i est $\mathbb{P}_\theta$.

Dans l'exemple introductif, faire des calculs sous $\mathbb{P}_\theta$ revient donc à faire des calculs en supposant que le biais de la pièce vaut θ. Par la suite, l'opérateur d'espérance associé à $\mathbb{P}_\theta$ sera noté $\mathbb{E}_\theta$.

Exemple 5.2.2. Considérons le modèle gaussien à variance connue $\{\mathbb{P}_m : m \in \mathbb{R}\}$ avec $\mathbb{P}_m$ la loi normale de moyenne m et de variance 1. On a alors par exemple $\mathbb{E}_1(X_1) = 1$, $\mathbb{E}_2(X_1) = 2$ et plus généralement $\mathbb{E}_m(X_1) = m$: cette dernière égalité est à interpréter de la manière suivante : si X_1 suit une loi normale de paramètre $(m, 1)$, alors sa moyenne vaut m.

Par la suite, on supposera par simplicité que les X_i sont à valeurs dans $\mathbb{R}$ (la généralisation au cas vectoriel ne posant pas de problème technique majeur), et on notera les vecteurs en gras et notamment $\mathbf{X}_n = (X_1, \ldots, X_n)$ et $\mathbf{x}_n \in \mathbb{R}^n$ une réalisation possible de $\mathbf{X}_n$. Enfin, supposera que pour tout $\theta \in \Theta$, le second moment de X_1 sous $\mathbb{P}_\theta$ est fini.

Définition 5.2.1. On appelle statistique toute fonction de $(X_1, \ldots, X_n)$ à valeurs réelle ou vectorielle indépendante de θ.

Exemple 5.2.3. Considérons le problème d'estimer la variance σ^* d'un échantillon tiré selon une loi normale de moyenne μ^* et de variance σ^*. Si l'on suppose la moyenne μ^* connue,

alors

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu^*)^2 \tag{5.1}$$

est bien une statistique, qui plus est naturelle puisqu'elle converge presque sûrement vers la variance $(\sigma^*)^2$ par la loi des grands nombres. En revanche, si la moyenne μ^* est inconnue alors cette variable aléatoire n'est plus une statistique puisqu'elle fait appel au paramètre inconnu μ^*.

Définition 5.2.2. On appelle estimateur toute statistique à valeurs dans Θ.

Cette définition n'est évidemment pas très intéressante puisqu'extrêmement large : ce qui rendra un estimateur intéressant sont les propriétés qu'il satisfait.

Définition 5.2.3. Un estimateur $\hat{\theta}$ est sans biais si $\mathbb{E}_\theta(\hat{\theta}) = \theta$ pour tout $\theta \in \Theta$, et une suite d'estimateurs $(\hat{\theta}_n, n \in \mathbb{N}^*)$ est asymptotiquement sans biais si $\mathbb{E}_\theta(\hat{\theta}_n) \to \theta$ pour tout $\theta \in \Theta$.

Définition 5.2.4. Une suite d'estimateurs $(\hat{\theta}_n, n \in \mathbb{N}^*)$ est dite convergente si pour tout $\theta \in \Theta$, $\hat{\theta}_n \overset{\text{p.s.}}{\to} \theta$ sous $\mathbb{P}_\theta$.

Exemple 5.2.4. On considère l'exemple du modèle gaussien $\{\mathbb{P}_\sigma : \sigma \in \mathbb{R}_+\}$ à moyenne μ^* connue, i.e., $\mathbb{P}_\sigma$ est la loi normale de moyenne μ^* et de variance σ^2 : ainsi, $\hat{\sigma}_n^2$ défini par (5.1) est un estimateur sans biais et convergent de σ^2.

Puisqu'un estimateur n'est rien d'autre qu'une variable aléatoire, il existe autant de mode de convergence pour les estimateurs que pour les variables aléatoires. Par exemple, une suite d'estimateurs est :

Convergente en loi si $\hat{\theta}_n \overset{\text{L}}{\to} \theta$ sous $\mathbb{P}_\theta$ pour tout $\theta \in \Theta$;

Convergente en probabilité si $\hat{\theta}_n \overset{\mathbb{P}}{\to} \theta$ sous $\mathbb{P}_\theta$ pour tout $\theta \in \Theta$;

Convergente en moyenne quadratique si $\mathbb{E}_\theta[(\hat{\theta}_n - \theta)^2] \to 0$ pour tout $\theta \in \Theta$.

5.2.2 Existence d'une densité

Dans le reste du cours on supposera que, pour tout $\theta \in \Theta$, $\mathbb{P}_\theta$ est soit discrète, soit absolument continue : dans le second cas, on notera alors f_θ sa densité. Pour $x \in \mathbb{R}$ et $\theta \in \Theta$ on définit alors $p(x;\theta)$ de la manière suivante :

$$p(x;\theta) = \begin{cases} \mathbb{P}_\theta(X = x) & \text{dans le cas discret,} \\ f_\theta(x) & \text{dans le cas continu,} \end{cases}$$

et on appellera $p(\,\cdot\,;\theta)$ la densité de $\mathbb{P}_\theta$ (dans le cas absolument continu c'est cohérent avec ce qu'on a fait au Chapitre 2, et dans le cas discret la théorie de la mesure permettrait de justifier cette terminologie). Pour $\mathbf{x}_n = (x_1, \ldots, x_n) \in \mathbb{R}^n$ on notera

$$p_n(\mathbf{x}_n;\theta) = \prod_{k=1}^{n} p(x_k;\theta)$$

qui correspond à la densité du vecteur $(X_1, \ldots, X_n)$ sous $\mathbb{P}_\theta$, puisque que les X_k sous sont supposés i.i.d. sous chaque $\mathbb{P}_\theta$.

Lorsque nous aurons des calculs à effectuer nous supposerons par commodité que les mesures de probabilités impliquées sont absolument continues, mais des calculs similaires (en remplaçant intégrale par somme et densité par loi) marchent dans le cas discret.

Remarque 5.2.1. Dans ce cours on se limite principalement au cas de la dimension 1 : néanmoins, tous les résultats et preuves ci-dessous se généralisent aisément à toute dimension, essentiellement en remplaçant la dérivée première ∂_θ par le gradient ∇_θ et la dérivée seconde ∂^2_θ par la matrice hessienne ∇^2_θ.

5.3 Vraisemblance : l'estimation paramétrique comme un problème inverse de probabilités

Le problème d'estimation paramétrique tel qu'on l'a présenté ci-dessus est d'une certaine manière le dual des questions de probabilité qu'on s'est posées dans la première partie du cours. On peut caricaturer les deux approches de la manière suivante :

Point de vue probabiliste : étant donné une loi de probabilité $\mathbb{P}$, des variables aléatoires générées selon $\mathbb{P}$ vont-elles satisfaire certaines propriétés de régularité malgré le hasard inhérent ? Par exemple, la loi des grands nombres qui assure que, sous $\mathbb{P}$, la moyenne empirique converge vers la moyenne met un peu d'ordre dans le chaos.

Point de vue statistique : étant donné des observations $X_1, \dots, X_n$, quelle loi de probabilité explique le mieux ces données ?

Ainsi, si en probabilités on part de $\mathbb{P}$ et on s'intéresse aux propriétés satisfaites par une suite i.i.d., en statistique on effectue la démarche inverse : la suite $X_1, \dots, X_n$ est donnée, ce sont les observations dont on va se servir pour retrouver $\mathbb{P}$. Ce changement radical de point de vue justifie d'utiliser une terminologie différente comme dans le cas de la vraisemblance.

Définition 5.3.1. La **vraisemblance** $\mathcal{L}$ est la fonction définie par

$$\mathcal{L} : (\theta, \mathbf{x}) \in \Theta \times \mathbb{R}^n \mapsto \mathcal{L}(\theta; \mathbf{x}) = p_n(\mathbf{x}; \theta).$$

On s'accommodera du léger abus de notation du fait que la vraisemblance dépend de la taille n de l'échantillon. On notera par ailleurs que, pour θ fixé, la vraisemblance n'est rien d'autre que la densité de $\mathbf{X}_n$ sous $\mathbb{P}_\theta$.

Cette terminologie traduit le changement de point de vue entre probabilités et statistique : en probabilités, le paramètre θ est fixe et l'on s'intéresse à la probabilité d'obtenir un résultat donné, l'accent est donc mis sur les réalisations $\mathbf{x}$ possibles. En statistique en revanche, le point de vue est renversé : les données sont ce qu'elles sont, par contre on peut jouer sur le paramètre de la loi de probabilité pour voir son influence sur la probabilité d'observer ces données. Ainsi,

$\mathcal{L}(\theta; \mathbf{x})$ **est la vraisemblance que le paramètre** θ **explique les observations** $\mathbf{x}$**.**

5.4 Modèle régulier, vecteur du score et information de Fisher

Pour $g : (x, \theta) \in \mathbb{R} \times \Theta \mapsto g(x; \theta) \in \mathbb{R}$ on note par la suite $\partial_\theta g(x; \theta)$ et $\partial^2_\theta g(x; \theta)$ ses dérivées partielle du premier et du second ordre en θ, à x fixé, lorsqu'elles sont bien définies. On réunit ici plusieurs hypothèses qui seront fréquemment invoquées :

(H1) le support des lois $\mathbb{P}_\theta$ ne dépend pas de θ, i.e., l'ensemble $\{x \in \mathbb{R} : p(x; \theta) > 0\}$ ne dépend pas de θ ;

(H2) pour tout $x \in \mathbb{R}$, $\theta \in \Theta \mapsto p(x;\theta)$ est deux fois différentiable ;

(H3) dans tous les cas rencontrés, on peut intervertir intégrable (ou somme) et dérivation.

Il s'ensuit par exemple de cette dernière hypothèse que, dans le cas absolument continu,

$$\int \partial_\theta p(x;\theta)\mathrm{d}x = 0. \tag{5.2}$$

En effet, puisque l'on peut intervertir intégrale et dérivation, on a

$$\int \partial_\theta p(x;\theta)\mathrm{d}x = \partial_\theta \int p(x;\theta)\mathrm{d}x$$

et puisque $p(\,\cdot\,;\theta)$ à θ fixé est une densité de probabilité, la fonction $\theta \mapsto \int p(x;\theta)\mathrm{d}x$ est constante (et prend la valeur 1) et sa dérivée est donc nulle.

Définition 5.4.1. Lorsque l'hypothèse (H2) est satisfaite, le vecteur $V_n(\theta) = \partial_\theta \ln p_n(\mathbf{X}_n;\theta)$ est appelé **vecteur du score** de $\mathbf{X}_n$.

On considère alors une quatrième hypothèse :

(H4) pour tout $\theta \in \Theta$, $V_1(\theta)$ est de carré intégrable sous $\mathbb{P}_\theta$.

Définition 5.4.2. Si Θ est un ouvert et que les quatre hypothèses (H1)–(H4) sont satisfaites, le modèle $\mathcal{P}$ est dit **régulier**.

Lorsque le modèle est régulier, la variance du vecteur du score est bien définie : cette variance est appelée information de Fisher. L'information de Fisher apparaît dans la borne de Fréchet–Darmois–Cramer–Rao (Théorème 5.6.1 ci-dessous) et joue plus généralement un rôle important en théorie de l'information.

Définition 5.4.3. Lorsque le modèle est régulier, la variance de $V_1(\theta)$ est appelée **information de Fisher** et est notée $I(\theta) = \mathbb{V}\mathrm{ar}_\theta\left(V_1(\theta)\right)$.

Proposition 5.4.1. *Si le modèle est régulier, alors l'information de Fisher est donnée par*

$$I(\theta) = -\mathbb{E}_\theta\left(\partial_\theta^2 \ln p(X_1;\theta)\right).$$

Démonstration. On montre d'abord que $V_1(\theta)$ est centrée. On a

$$\mathbb{E}_\theta\left(V_1(\theta)\right) = \mathbb{E}_\theta\left(\partial_\theta \ln p(X_1;\theta)\right) = \int p(x;\theta)\partial_\theta \ln p(x;\theta)\mathrm{d}x = \int \partial_\theta p(x;\theta)\mathrm{d}x = 0$$

où l'avant-dernière égalité vient par intégration par parties, et la dernière égalité suit de (5.2). Cela montre que $V_1(\theta)$ est centrée, et on montre maintenant la formule pour l'information de Fisher. Puisque $\partial_\theta p = p\partial_\theta \ln p$, (5.2) se réécrit

$$0 = \int \partial_\theta \ln p(x;\theta)p(x;\theta)\mathrm{d}x$$

et on obtient donc en dérivant par rapport à θ

$$0 = \int \partial_\theta^2 \ln p(x;\theta)p(x;\theta)\mathrm{d}x + \int \left(\partial_\theta \ln p(x;\theta)\right)^2 p(x;\theta)\mathrm{d}x.$$

Le premier terme du membre de droite est égal à $\mathbb{E}(\partial_\theta^2 \ln p(X_1;\theta))$ et le second terme à $\mathbb{E}((\partial_\theta \ln p(X_1;\theta))^2) = \mathbb{E}(V_1(\theta)^2) = \mathbb{V}\mathrm{ar}(V_1(\theta)) = I(\theta)$, ce qui prouve bien le résultat. ∎

5.5 Estimateurs classiques

Chaque problème d'estimation paramétrique, lié à un modèle paramétrique donné, doit en général se résoudre de manière ad hoc. Il existe néanmoins deux grandes familles d'estimateurs universels que nous présentons maintenant : l'estimateur du maximum de vraisemblance et l'estimateur d'une moyenne.

5.5.1 Estimateur du maximum de vraisemblance

On considère le modèle paramétrique $\{\mathbb{P}_{0,1}, \mathbb{P}_{0,9}\}$ où $\mathbb{P}_p$ est la loi de Bernoulli de paramètre $p \in [0,1]$: en d'autres termes, on cherche à estimer le biais d'une pièce sachant que le biais (= probabilité d'obtenir pile) vaut soit 0,1, soit 0,9. Si l'on ne dispose du résultat que d'un seul lancer et que l'on a observé pile, quelle est la meilleure décision à prendre ?

- Si le biais valait 0,1, la probabilité d'obtenir pile aurait été de 0,1 ;
- Si le biais valait 0,9, la probabilité d'obtenir pile aurait été de 0,9.

Intuitivement, on opte donc pour la deuxième solution qui explique mieux le résultat obtenu : si l'on a obtenu pile, c'est probablement que c'était un résultat plus probable. De manière un peu plus formel, le choix effectué est revenu à

Sélectionner le paramètre qui maximise la probabilité des observations.

Cette approche très naturelle est à la base de l'estimation par maximum de vraisemblance.

Définition 5.5.1. On suppose que pour $\mathbf{x} \in \mathbb{R}^n$ fixé, la fonction $\theta \in \Theta \mapsto \mathcal{L}(\theta; \mathbf{x})$ admet un unique maximum. Alors l'estimateur

$$\hat{\theta}_n = \arg \max_{\theta \in \Theta} \mathcal{L}(\theta; \mathbf{X}_n)$$

est appelé l'**estimateur du maximum de vraisemblance de** θ^*.

Le résultat suivant est un résultat préliminaire qui sera complété avec le Théorème 5.6.2 lorsque nous aurons vu la notion d'efficacité.

Théorème 5.5.1. *Si le modèle est régulier, alors l'estimateur du maximum de vraisemblance est convergent en loi.*

Pour calculer l'estimateur du maximum de vraisemblance en pratique, on utilisera le fait que $\mathcal{L}(\theta; \mathbf{x}) = \prod_{k=1}^n \mathcal{L}(\theta; x_k)$ par indépendance. A cause de cette forme produit, il sera aussi souvent plus facile de maximiser le logarithme $\ln \mathcal{L}$ de la vraisemblance – appelé log-vraisemblance – plutôt que la vraisemblance elle-même : puisque la fonction ln est croissante ces deux problèmes d'optimisation sont équivalents.

Exemple 5.5.1. Considérons le modèle paramétrique gaussien $\{\mathbb{P}_m : m \in \mathbb{R}\}$ avec $\mathbb{P}_m$ la loi normale de moyenne m et de variance σ^2 connue. Pour $\mathbf{x} \in \mathbb{R}^n$ on a alors

$$\ln \mathcal{L}(m; \mathbf{x}) = \ln \prod_{k=1}^n \frac{1}{\sigma(2\pi)^{1/2}} \exp\left(-\frac{(x_k - m)^2}{2\sigma^2}\right) = -\frac{1}{2\sigma^2} \sum_{k=1}^n (x_k - m)^2 - \frac{n}{2} \ln(2\pi\sigma^2)$$

et donc

$$\frac{\partial \mathcal{L}}{\partial m}(m;\mathbf{x}) = 0 \iff -\frac{1}{\sigma^2}\sum_{k=1}^{n}(m - x_k) = 0 \iff m = \frac{1}{n}\sum_{k=1}^{n} x_k.$$

Ainsi, l'estimateur du maximum de vraisemblance est donné par

$$\hat{\theta}_n = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

Exemple 5.5.2. Considérons le modèle paramétrique exponentiel $\{\mathbb{P}_\lambda : \lambda \in \mathbb{R}_+\}$ avec $\mathbb{P}_\lambda$ la loi exponentielle de paramètre λ. Pour $\mathbf{x} \in \mathbb{R}_+^n$ on a alors

$$\ln \mathcal{L}(\lambda;\mathbf{x}) = \sum_{k=1}^{n} \ln\left(\lambda e^{-\lambda x_k}\right) = n\ln\lambda - \lambda\sum_{k=1}^{n} x_k$$

et donc

$$\frac{\partial \mathcal{L}}{\partial \lambda}(\lambda;\mathbf{x}) = 0 \iff \frac{n}{\lambda} - \sum_{k=1}^{n} x_k = 0 \iff \lambda = \frac{n}{\sum_{k=1}^{n} x_k}.$$

Ainsi, l'estimateur du maximum de vraisemblance est donné par

$$\hat{\theta}_n = \frac{n}{\sum_{k=1}^{n} X_k}.$$

Remarque 5.5.1. Il est très important de ne pas confondre $\mathbf{x} \in \mathbb{R}^n$, un nombre déterministe qui représente une réalisation potentielle de la variable aléatoire $\mathbf{X}_n$, et la variable aléatoire $\mathbf{X}_n$ elle-même. Ainsi, dans l'exemple ci-dessus on a fait les calculs avec $\mathbf{x}$ et l'on a trouvé comme maximiseur $(1/n)\sum_{k=1}^{n} x_k$, mais l'estimateur du maximum de vraisemblance est bien $(1/n)\sum_{k=1}^{n} X_k$ et non $(1/n)\sum_{k=1}^{n} x_k$.

5.5.2 Estimateur de la moyenne (et donc de la variance)

On présente maintenant des estimateurs très naturels, dont la justification repose sur la loi forte des grands nombres.

5.5.2.1 Estimateur de la moyenne

Si l'on veut estimer la moyenne de $\mathbb{P}$, un estimateur très naturel est donné par la moyenne empirique

$$\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

La linéarité de l'espérance nous assure que $\bar{X}_n$ est sans biais, et la loi forte des grands nombres garantit que $\bar{X}_n$ est un estimateur convergent de la moyenne. En outre, la moyenne empirique est parfois égale à l'estimateur du maximum de vraisemblance comme le montre l'exemple 5.5.1.

5.5.2.2 Estimateur de la variance à moyenne connue

Par définition, la variance de X est donnée par $\mathbb{V}\mathrm{ar}(X) = \mathbb{E}((X-m)^2)$ avec $m = \mathbb{E}(X)$: ainsi, estimer la variance revient à estimer la moyenne de la variable aléatoire $(X-m)^2$ et si la moyenne m est connue, on peut donc appliquer la méthode ci-dessus.

5.5.2.3 Estimateur de la variance à moyenne inconnue

En revanche, si la moyenne est inconnue l'estimateur de la variance empirique

$$\frac{1}{n}\sum_{k=1}^{n}(X_k - m)^2$$

n'est plus acceptable. L'idée est alors de remplacer la moyenne m par son estimation $\bar{X}_n$, ce qui nous mène à l'estimateur suivant :

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \bar{X}_n\right)^2.$$

La justification de la division par $n-1$ réside dans le résultat suivant.

Proposition 5.5.2. *S_{n-1}^2 est un estimateur sans biais de la variance.*

Démonstration. Il s'agit de montrer que $\mathbb{E}(S_{n-1}^2) = \mathbb{V}\mathrm{ar}(X_1)$ et puisque

$$\mathbb{E}\left(S_{n-1}^2\right) = \mathbb{E}\left(\frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2\right) = \frac{n}{n-1}\mathbb{V}\mathrm{ar}\left(X_1 - \bar{X}_n\right)$$

(en utilisant par symétrie que les $X_k - \bar{X}_n$ ont même loi [1]) il suffit de montrer l'identité $\mathbb{V}\mathrm{ar}(X_1 - \bar{X}_n) = (n-1)\mathbb{V}\mathrm{ar}(X_1)/n$. Par définition de $\bar{X}_n$ on a

$$\mathbb{V}\mathrm{ar}\left(X_1 - \bar{X}_n\right) = \mathbb{V}\mathrm{ar}\left(\frac{n-1}{n}X_1 - \frac{1}{n}\sum_{k=2}^{n}X_k\right) = \frac{(n-1)^2}{n^2}\mathbb{V}\mathrm{ar}(X_1) + \frac{1}{n^2}\mathbb{V}\mathrm{ar}\left(\sum_{k=2}^{n}X_k\right)$$

en utilisant pour la dernière égalité le fait que X_1 et $\sum_{k=2}^{n}X_k$ sont indépendantes, et puisque $\mathbb{V}\mathrm{ar}(X_2 + \cdots + X_n) = (n-1)\mathbb{V}\mathrm{ar}(X_1)$ puisque les variables $X_2, \ldots, X_n$ sont i.i.d. on obtient bien le résultat. ■

Proposition 5.5.3. *S_{n-1}^2 est un estimateur convergent de la variance.*

Démonstration. On calcule

$$S_{n-1}^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{k=1}^{n}X_k^2 - \bar{X}_n^2\right)$$

et le résultat suit donc d'une double application de la loi forte des grands nombres, qui donne $\frac{1}{n}\sum_{k=1}^{n}X_k^2 \overset{\text{p.s.}}{\to} \mathbb{E}(X_1^2)$ et $\bar{X}_n^2 \overset{\text{p.s.}}{\to} \mathbb{E}(X_1)^2$. ■

5.6 Estimateurs efficaces

Dans cette section on cherche à quantifier la performance d'un estimateur. La borne de Fréchet–Darmois–Cramer–Rao établit une borne inférieure générale qui mène alors naturellement à la définition d'efficacité : un estimateur est efficace s'il atteint cette borne.

1. Attention : ces variables ne sont pas indépendantes à cause de la corrélation induite par $\bar{X}_n$!

5.6.1 De l'importance de la variance

La variance quantifie la vitesse à laquelle l'estimateur $\hat{\theta}_n$ converge. D'une part, la Proposition 3.3.1 montre que, pour un estimateur sans biais, il suffit de montrer que sa variance tend vers 0 pour garantir que l'estimateur converge en probabilités.

Exemple 5.6.1. On considère le modèle paramétrique exponentiel où l'on choisit comme paramètre la moyenne de la loi exponentielle (et non son paramètre comme dans l'exemple 5.5.2) :

$$\Theta =]0, \infty[\text{ et } \mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\} \text{ avec } \mathbb{P}_\theta \text{ la loi exponentielle de paramètre } \theta^{-1}.$$

Puisque $\mathbb{E}_\theta(X_1) = \theta$ et $\mathbb{V}\mathrm{ar}_\theta(X_1) = \theta^2$, on a

$$\mathbb{E}_\theta(\hat{\theta}_n) = \mathbb{E}_\theta(X_1) = \theta \quad \text{et} \quad \mathbb{V}\mathrm{ar}_\theta(\hat{\theta}_n) = \frac{1}{n}\mathbb{V}\mathrm{ar}_\theta(X_1) = \frac{\theta^2}{n} \underset{n\to\infty}{\longrightarrow} 0$$

et donc la Proposition 3.3.1 implique que $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$ (sous $\mathbb{P}_\theta$, pour tout $\theta \in \Theta$).

De manière plus fine, la vitesse à laquelle la variance tend vers 0 quantifie la vitesse à laquelle $\hat{\theta}_n$ converge. Pour formaliser cela, on considère la distance en moyenne quadratique, donnée par la norme $L_2(\Omega, \mathbb{P}_\theta)$: pour une variable aléatoire Y,

$$\|Y\|_{2,\theta} = \sqrt{\mathbb{E}_\theta\left(Y^2\right)}.$$

Cette distance contrôle la vitesse de convergence de $\hat{\theta}_n$ vers θ via l'inégalité de Bienaymé–Tchebychev (Théorème 1.3.12) qui nous dit que

$$\mathbb{P}_\theta\left(\|\hat{\theta}_n - \theta\|_{2,\theta} \geq \varepsilon\right) \leq \frac{\|\hat{\theta}_n - \theta\|_{2,\theta}^2}{\varepsilon^2}.$$

Dans le cas d'un estimateur sans biais, cette distance est égale à la variance $\hat{\theta}_n$:

$$\|\hat{\theta}_n - \theta\|_{2,\theta}^2 = \mathbb{E}_\theta\left[\left(\hat{\theta}_n - \theta\right)^2\right] = \mathbb{V}\mathrm{ar}(\hat{\theta}_n)$$

et l'inégalité précédente se réécrit donc

$$\mathbb{P}_\theta\left(\|\hat{\theta}_n - \theta\|_{2,\theta} \geq \varepsilon\right) \leq \frac{\mathbb{V}\mathrm{ar}(\hat{\theta}_n)}{\varepsilon^2}.$$

Ainsi, plus la variance tend vite vers 0 et plus la probabilité $\mathbb{P}_\theta\left(\|\hat{\theta}_n - \theta\|_{2,\theta} \geq \varepsilon\right)$ tend vite vers 0.

Parmi tous les estimateurs sans biais de θ on cherchera ceux de plus petite variance.

Si l'on autorise un biais, alors on a

$$\mathbb{E}_\theta[(\hat{\theta}_n - \theta)^2] = \mathbb{E}_\theta[(\hat{\theta}_n - \mathbb{E}_\theta(\hat{\theta}_n) + \mathbb{E}_\theta(\hat{\theta}_n) - \theta)^2] = \mathbb{V}\mathrm{ar}_\theta(\hat{\theta}_n) + \left[\mathbb{E}_\theta(\hat{\theta}_n) - \theta\right]^2$$

et la vitesse de convergence dépend à la fois de la vitesse à laquelle $\mathbb{V}\mathrm{ar}_\theta(\hat{\theta}_n) \to 0$, mais aussi à laquelle $\mathbb{E}_\theta(\hat{\theta}_n) \to \theta$.

5.6.2 Borne de Fréchet–Darmois–Cramer–Rao

Le Théorème 5.6.1 ci-dessous donne une borne inférieure générale sur la variance d'un estimateur sans biais, appelée **borne de Fréchet–Darmois–Cramer–Rao**, qui permet donc d'évaluer la vitesse de convergence d'un estimateur.

Théorème 5.6.1 (Borne de Fréchet–Darmois–Cramer–Rao)**.** *Si le modèle est régulier, que $I(\theta) > 0$ et que $\hat{\theta}_n$ est un estimateur de carré intégrable et sans biais de θ, alors*

$$\mathbb{V}\mathrm{ar}_\theta(\hat{\theta}_n) \geq \frac{1}{n} I(\theta)^{-1}.$$

Démonstration. Puisque $\hat{\theta}_n$ est sans biais, on a $\theta = \mathbb{E}_\theta(\hat{\theta}_n) = \int \hat{\theta}_n(\mathbf{x}) p_n(\mathbf{x};\theta) \mathrm{d}\mathbf{x}$ et en dérivant par rapport à θ, il vient donc

$$1 = \int \hat{\theta}_n(\mathbf{x}) \partial_\theta p_n(\mathbf{x};\theta) \mathrm{d}\mathbf{x}.$$

Puisque $\partial_\theta p_n = p_n \partial_\theta \ln p_n$, l'égalité précédente se réécrit $1 = \mathbb{E}_\theta(\hat{\theta}_n V_n(\theta))$. Puisque les X_k sont indépendantes, on remarque que

$$V_n(\theta) = \sum_{k=1}^{n} \partial_\theta \ln p(X_k;\theta)$$

et $V_n(\theta)$ est donc centré puisque $V_1(\theta)$ l'est. En particulier,

$$1 = \mathbb{E}_\theta\left(\hat{\theta}_n V_n(\theta)\right) = \mathbb{E}_\theta\left((\hat{\theta}_n - \theta) V_n(\theta)\right)$$

et donc l'inégalité de Cauchy–Schwarz (Théorème 1.3.13) donne

$$1 \leq \mathbb{V}\mathrm{ar}_\theta(\hat{\theta}_n)\mathbb{V}\mathrm{ar}_\theta(V_n(\theta)).$$

On conclut finalement en notant que $V_n(\theta)$ est égale à la somme de n variables i.i.d. distribuées comme $V_1(\theta)$, et donc $\mathbb{V}\mathrm{ar}_\theta(V_n(\theta)) = n\mathbb{V}\mathrm{ar}_\theta(V_1(\theta)) = nI(\theta)$. ■

Remarque 5.6.1. Dans le cas général $d \geq 1$, il faut remplacer l'hypothèse $I(\theta) > 0$ par l'hypothèse $I(\theta)$ inversible.

Cette borne donne lieu à la définition d'efficacité.

Définition 5.6.1. Un estimateur sans biais $\hat{\theta}_n$ est **efficace** si $\mathbb{V}\mathrm{ar}_\theta(\hat{\theta}_n) = (nI(\theta))^{-1}$, et **asymptotiquement efficace** si $n\mathbb{V}\mathrm{ar}_\theta(\hat{\theta}_n) \to I(\theta)^{-1}$.

Exemple 5.6.2. On continue l'exemple 5.6.1 du modèle paramétrique exponentiel – paramétré par la moyenne – pour lequel $p(x;\theta) = \theta^{-1} e^{-x/\theta} \mathbb{1}\{x \geq 0\}$. Ainsi, $\ln p(x;\theta) = -\ln\theta - x/\theta$ ce qui donne $\partial_\theta^2 \ln p(x;\theta) = \theta^{-2} - 2x\theta^{-3}$ et donc

$$I(\theta) = -\mathbb{E}_\theta\left(\partial_\theta^2 \ln p(X_1;\theta)\right) = \frac{2\mathbb{E}_\theta(X_1)}{\theta} - \frac{1}{\theta^2} = \frac{1}{\theta^2}.$$

Ainsi la borne de Fréchet–Darmois–Cramer–Rao vaut $1/(nI(\theta)) = \theta^2/n$ et puisque $\mathbb{V}\mathrm{ar}_\theta(\hat{\theta}_n) = \theta^2/n$ on en déduit que $\hat{\theta}_n$ est un estimateur efficace de θ.

5.6.3 Normalité asymptotique de l'estimateur du maximum de vraisemblance

Le résultat suivant complète le Théorème 5.5.1 et montre que l'estimateur du maximum de vraisemblance est asymptotiquement normal.

Théorème 5.6.2. *Si le modèle est identifiable et régulier, alors pour tout $\theta \in \Theta$ on a, sous $\mathbb{P}_\theta$,*

$$n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{\mathrm{L}} X$$

où X suit une loi normale centrée de variance $I(\theta)^{-1}$.

5.6.4 Réduction de la variance et statistique exhaustive

La borne de Fréchet–Darmois–Cramer–Rao fournit une indication de la performance d'un estimateur : une question naturelle est de savoir si cette borne est toujours atteignable. Plus généralement, on s'intéresse dans cette section à la question suivante : comment réduire la variance d'un estimateur convergent ? La réponse à cette question repose fondamentalement sur la notion de statistique exhaustive.

Définition 5.6.2. Une statistique T est dite **exhaustive** si la loi de X_1 sous $\mathbb{P}_\theta(\,\cdot \mid T)$ ne dépend pas de θ.

De manière un peu plus imagée,

> **T est une statistique exhaustive si elle contient toute l'information sur θ.**

L'intérêt d'une statistique exhaustive réside dans le résultat suivant qui montre que pour tout estimateur $\hat{\theta}$, alors $\mathbb{E}_\theta(\hat{\theta} \mid T)$ est un estimateur de variance plus faible. En fait, le fait que T soit exhaustive est nécessaire pour que $\mathbb{E}_\theta(\hat{\theta} \mid T)$ soit un estimateur : autrement, cette statistique dépendrait de θ !

En particulier, on omettra θ de cette notation et on écrira simplement $\mathbb{E}(\hat{\theta} \mid T)$. On remarque aussi immédiatement au vu du théorème de l'espérance totale que $\hat{\theta}$ et $\mathbb{E}(\hat{\theta} \mid T)$ ont le même biais, et le résultat suivant – qui n'est en fait qu'un cas particulier de l'inégalité de Jensen – montre que la variance est améliorée en considérant $\mathbb{E}(\hat{\theta} \mid T)$ à la place de $\hat{\theta}$.

Théorème 5.6.3 (Théorème de Rao–Blackwell)**.** *Si T est une statistique exhausive, alors pour tout estimateur $\hat{\theta}$ on a*

$$\mathbb{E}_\theta\left[\left(\mathbb{E}(\hat{\theta} \mid T) - \theta\right)^2\right] \leq \mathbb{E}_\theta\left[\left(\hat{\theta} - \theta\right)^2\right].$$

Démonstration. En considérant $\hat{\theta} - \theta$, on peut supposer sans perte de généralité que $\theta = 0$ et il faut donc montrer que $\mathbb{E}_\theta\left(\hat{\theta}^2 - (\mathbb{E}(\hat{\theta} \mid T))^2\right) \geq 0$. La Proposition 1.7.4 donne

$$\mathbb{E}(\theta\mathbb{E}(\theta \mid T)) = \mathbb{E}(\mathbb{E}(\theta \mid T)^2)$$

et donc

$$\mathbb{E}_\theta\left(\hat{\theta}^2 - (\mathbb{E}(\hat{\theta} \mid T))^2\right) = \mathbb{E}_\theta\left[\left(\hat{\theta} - \mathbb{E}(\hat{\theta} \mid T)\right)^2\right]$$

ce qui prouve le résultat. ∎

5.7 Régions de confiance

5.7.1 Généralités

Dans les sections précédentes on a construit des estimateurs $\hat{\theta}_n$ d'un paramètre θ^* inconnu. Néanmoins, dans la plupart des cas $\hat{\theta}_n$ n'est qu'une approximation de θ^* : dans le cas absolument continu par exemple on a $\mathbb{P}(\hat{\theta}_n = \theta^*) = 0$, et il se pose donc la question de savoir si cette approximation est bonne.

Pour répondre à cette question, on cherche, basé sur les observations $X_1, \dots, X_n$, un **ensemble aléatoire** $\Lambda_n \subset \Theta$ qui contienne le vrai paramètre θ^* avec grande probabilité (évidemment, sous la vraie mesure de probabilité). En d'autres termes, on cherche Λ_n tel que $\mathbb{P}(\Lambda_n \ni \theta^*)$ soit le plus large possible : on préfèrera noter $\mathbb{P}(\Lambda_n \ni \theta^*)$ plutôt que $\mathbb{P}(\theta^* \in \Lambda_n)$ pour mettre en avant que l'aléa porte sur Λ_n et non sur θ^*. Une généralisation de cette question consiste à chercher, pour chaque $\theta \in \Theta$, Λ_n tel que $\mathbb{P}_\theta(\Lambda_n \ni \theta)$ soit large.

Définition 5.7.1. Soit $\alpha \in [0, 1]$. L'ensemble $\Lambda_\alpha \subset \Theta$ est une **région de confiance de niveau** $1 - \alpha$ si $\mathbb{P}_\theta(\Lambda_\alpha \ni \theta) \geq 1 - \alpha$ pour tout $\theta \in \Theta$.

L'idée de base pour construire des intervalles de confiance, que l'on illustrera sur plusieurs exemples ci-dessous, peut être résumée de la manière suivante :

Pour construire des intervalles de confiance, on utilise des statistiques ayant des lois connues et indépendantes du paramètre à estimer

On illustre cette idée dans la section suivante dans le cas du modèle paramétrique gaussien : dans ce cas, les statistiques utilisées sont les moyennes et variances empiriques, dont les lois sont données par le Théorème 4.6.2.

5.7.2 Intervalles de confiance pour le modèle gaussien

5.7.2.1 Intervalles de confiance pour la moyenne à variance connue

On fixe $\sigma \in]0, \infty[$ et on considère le modèle paramétrique $\{\mathbb{P}_m : m \in \mathbb{R}\}$ avec $\mathbb{P}_m$ la loi normale de moyenne m et de variance σ^2. Alors $\bar{X}_n = n^{-1}(X_1 + \cdots + X_n)$ sous $\mathbb{P}_m$ suit une loi normale de moyenne m et de variance σ^2/n, en particulier on peut vérifier que c'est un estimateur sans biais et efficace de m. Le problème de cette statistique est que sa loi dépend de la moyenne m que l'on cherche à estimer, ce qui empêche de calculer la probabilité que $\bar{X}_n$ appartienne à un intervalle donné.

Pour pallier ce problème et se ramener à une loi indépendante de m, il suffit juste de centrer la variable : en effet, sous $\mathbb{P}_m$ la variable aléatoire

$$\frac{n^{1/2}}{\sigma}(\bar{X}_n - m)$$

suit la loi normale standard qui ne dépend plus de m. En particulier, si F est la fonction de répartition de la loi normale standard, i.e.,

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} \mathrm{d}y,$$

alors pour tout $a^+, a^- > 0$ on a

$$\mathbb{P}_m\left(-a^- < \frac{n^{1/2}}{\sigma}(\bar{X}_n - m) < a^+\right) = F(a^+) - F(-a^-).$$

On remarquera que le membre de gauche, alors qu'il a l'air de dépendre de m, n'en dépend en fait pas. La relation ci-dessus entraîne le résultat suivant.

Proposition 5.7.1. *Soit F la fonction de répartition de la loi normale standard. Alors pour tout couple $a^+, a^- > 0$ satisfaisant*

$$F(a^+) - F(-a^-) = 1 - \alpha, \tag{5.3}$$

l'intervalle aléatoire

$$\left[\bar{X}_n - \frac{\sigma a^-}{n^{1/2}},\ \bar{X}_n + \frac{\sigma a^+}{n^{1/2}}\right]$$

est un intervalle de confiance de niveau $1 - \alpha$.

Démonstration. Pour montrer le résultat, il s'agit de montrer que

$$\mathbb{P}_m\left(\left[\bar{X}_n - \frac{\sigma a^-}{n^{1/2}},\ \bar{X}_n + \frac{\sigma a^+}{n^{1/2}}\right] \ni m\right) = 1 - \alpha.$$

Cette probabilité se réécrit

$$\mathbb{P}_m\left(\bar{X}_n - \frac{\sigma a^-}{n^{1/2}} < m < \bar{X}_n + \frac{\sigma a^+}{n^{1/2}}\right) = \mathbb{P}_m\left(-a^- < \frac{n^{1/2}(m - \bar{X}_n)}{\sigma} < a^+\right)$$

qui est égale à $F(a^+) - F(-a^-)$ puisque $n^{1/2}(m - \bar{X}_n)/\sigma$ suit une loi normale standard. ■

On notera que la condition (5.3) est équivalente à $F(a^+) + F(a^-) = 2 - \alpha$ du fait que $F(-a) = 1 - F(a)$ par symétrie de la loi normale standard. Le résultat précédent nous donne une infinité d'intervalles de confiance. On en distinguera trois particuliers :

L'intervalle de confiance bilatéral symétrique : il s'agit de l'intervalle de confiance mettant autant de masse à droite qu'à gauche de $\bar{X}_n$, i.e., tel que

$$\mathbb{P}_\theta\left(\bar{X}_n - \frac{\sigma a^-}{n^{1/2}} < \theta < \bar{X}_n\right) = \mathbb{P}_\theta\left(\bar{X}_n < \theta < \bar{X}_n + \frac{\sigma a^+}{n^{1/2}}\right) = \frac{1 - \alpha}{2}.$$

Cette contrainte donne alors $a^+ = a^-$ (par symétrie de la loi normale) et donc (en utilisant la condition $F(a^+) + F(a^-) = 2 - \alpha$, équivalente à (5.3) comme expliqué précédemment)

$$F(a^+) = \frac{2 - \alpha}{2} \iff a^+ = F^{-1}(1 - \alpha/2)$$

ce qui donne, pour l'intervalle de confiance bilatéral symétrique,

$$\left[\bar{X}_n - \frac{\sigma F^{-1}(1 - \alpha/2)}{n^{1/2}}, \bar{X}_n + \frac{\sigma F^{-1}(1 - \alpha/2)}{n^{1/2}}\right];$$

L'intervalle de confiance unilatéral à droite : il s'agit de l'intervalle de confiance mettant le moins de masse possible à gauche de $\bar{X}_n$. Pour $\alpha < 1/2$ (cas intéressant en pratique), il convient donc de prendre $a^+ = \infty$ (qui donne par symétrie de la loi normale une masse $1/2$ à droite de $\bar{X}_n$) puis a^- qui satisfait $F(a^-) = 1-\alpha$, ce qui donne finalement l'intervalle

$$\left[\bar{X}_n - \frac{\sigma F^{-1}(1-\alpha)}{n^{1/2}}, \infty\right[;$$

L'intervalle de confiance unilatéral à gauche : il s'agit de l'intervalle de confiance mettant le moins de masse possible à droite de $\bar{X}_n$, et est obtenu par un raisonnement symétrique par rapport à l'intervalle de confiance unilatéral à droite.

Pour construire les intervalles de confiance ci-dessus on a utilisé la fonction F^{-1}, l'inverse de la fonction $F : \mathbb{R} \to [0,1]$ qui est continue et strictement croissante : cet inverse apparaît naturellement au vu de la condition (5.3). Pour toute fonction de répartition continue et strictement croissante, cet inverse est bien défini et permet de définir les quantiles d'une loi de distribution.

Définition 5.7.2. Soit F la fonction de répartition d'une loi de probabilité telle que F est continue et strictement croissante. Pour $\beta \in (0,1)$, $F^{-1}(\beta)$ est appelé **quantile d'ordre** β de F.

Ainsi, par définition, le quantile d'ordre β est l'unique nombre γ tel que $F(\gamma) = \beta$.

5.7.2.2 Intervalle de confiance pour la variance à moyenne connue

On fixe maintenant $m \in \mathbb{R}$ et on considère le modèle paramétrique $\{\mathbb{P}_\sigma : \sigma \in \mathbb{R}_+\}$ avec $\mathbb{P}_\sigma$ la loi normale de moyenne m fixée et de variance σ^2. Comme dans le cas précédent, afin de construire un intervalle de confiance on cherche à se ramener à une loi connue. Puisque la moyenne est connue, l'estimateur naturel de la variance est donné par

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - m)^2.$$

La loi de cette variable aléatoire dépend de σ, mais puisque $(X_k - m)/\sigma$ suit une loi normale standard, on voit que $n\hat{\sigma}_n^2/\sigma^2$ sous $\mathbb{P}_\sigma$ suit une loi du χ^2 à n degrés de liberté et on obtient le résultat suivant.

Proposition 5.7.2. *Soit G_n la fonction de répartition de la loi du χ^2 à n degrés de liberté. Alors pour tout couple $a^+ > a^- > 0$ satisfaisant*

$$G_n(a^+) - G_n(a^-) = 1 - \alpha, \tag{5.4}$$

l'intervalle aléatoire

$$\left[\frac{n\hat{\sigma}_n^2}{a^+}, \frac{n\hat{\sigma}_n^2}{a^-}\right]$$

est un intervalle de confiance de niveau $1-\alpha$.

Démonstration. Pour montrer le résultat il faut montrer que

$$\mathbb{P}_\sigma\left(\left[\frac{n\hat{\sigma}_n^2}{a^+}, \frac{n\hat{\sigma}_n^2}{a^-}\right] \ni \sigma^2\right) = 1 - \alpha.$$

Pour établir cette égalité, on écrit

$$\mathbb{P}_\sigma\left(\left[\frac{n\hat{\sigma}_n^2}{a^+},\ \frac{n\hat{\sigma}_n^2}{a^-}\right] \ni \sigma^2\right) = \mathbb{P}_\sigma\left(\frac{n\hat{\sigma}_n^2}{a^+} < \sigma^2 < \frac{n\hat{\sigma}_n^2}{a^-}\right) = \mathbb{P}_\sigma\left(a^- < \frac{n\hat{\sigma}_n^2}{\sigma^2} < a^+\right)$$

et puisque $n\hat{\sigma}_n^2/\sigma^2$ sous $\mathbb{P}_\sigma$ suit une loi du χ^2 à n degrés de liberté, on a donc par définition de G_n

$$\mathbb{P}_\sigma\left(a^- < \frac{n\hat{\sigma}_n^2}{\sigma^2} < a^+\right) = G_n(a^+) - G_n(a^-)$$

ce qui donne le résultat. ■

Un choix classique pour a^+, a^- satisfaisant (5.4) est donné par

$$a^- = G_n^{-1}(\alpha/2) \quad \text{et} \quad a^+ = G_n^{-1}(1-\alpha/2),$$

qui correspond à imposer la condition symétrique supplémentaire $G_n(a^+) + G_n(a^-) = 1$.

5.7.2.3 Intervalle de confiance pour la variance à moyenne inconnue

On considère le même modèle paramétrique que précédemment mais on suppose maintenant que la moyenne est inconnue. Dans ce cas, l'estimateur de la variance est obtenu en remplaçant la moyenne par la moyenne empirique : il est donné par

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \bar{X}_n\right)^2$$

et $(n-1)S_{n-1}^2/\sigma^2$ suit une loi du χ^2 à $n-1$ degrés de liberté d'après le Théorème 4.6.2. Ainsi, en suivant exactement la même démarche que précédemment on obtient le résultat suivant. Comme dans la Proposition 5.7.2 G_n désigne la fonction de répartition du χ^2 à n degrés de liberté.

Proposition 5.7.3. *Pour tout couple $a^+ > a^- > 0$ satisfaisant*

$$G_{n-1}(a^+) - G_{n-1}(a^-) = 1 - \alpha,$$

l'intervalle aléatoire

$$\left[\frac{(n-1)S_{n-1}^2}{a^+},\ \frac{(n-1)S_{n-1}^2}{a^-}\right]$$

est un intervalle de confiance de niveau $1-\alpha$.

De même que précédemment, une condition symétrique revient à prendre $a^- = G_{n-1}^{-1}(\alpha/2)$ et $a^+ = G_{n-1}^{-1}(1-\alpha/2)$.

5.7.2.4 Intervalle de confiance pour la moyenne à variance inconnue

On conclut par la construction d'intervalles de confiance pour la moyenne lorsque la variance est inconnue : on considère donc le modèle paramétrique $\{\mathbb{P}_{m,\sigma} : m \in \mathbb{R}, \sigma \in \mathbb{R}_+\}$ avec $\mathbb{P}_{m,\sigma}$ la loi normale de moyenne m et de variance σ^2 (on sort donc du cadre de la dimension un considéré dans le reste du chapitre). Encore une fois, le but est de se ramener à une loi connue : ici, on utilisera le fait que

$$n^{1/2}\frac{\bar{X}_n - m}{S_{n-1}}$$

sous $\mathbb{P}_{m,\sigma}$ suit une loi de Student à $n-1$ degrés de liberté (cf. Théorème 4.6.2). Sans surprise, les intervalles de confiance feront donc appel à la fonction de répartition de la loi de Student.

Proposition 5.7.4. *Soit H_n la fonction de répartition de la loi de Student à n degrés de liberté. Alors pour tout couple $a^+, a^- > 0$ satisfaisant*

$$H_{n-1}(a^+) - H_{n-1}(-a^-) = 1 - \alpha,$$

l'intervalle aléatoire

$$\left[\bar{X}_n - \frac{S_{n-1}a^-}{n^{1/2}},\ \bar{X}_n + \frac{S_{n-1}a^+}{n^{1/2}}\right]$$

est un intervalle de confiance de niveau $1-\alpha$.

Démonstration. Pour montrer le résultat, il s'agit de montrer que

$$\mathbb{P}_{m,\sigma}\left(\left[\bar{X}_n - \frac{S_{n-1}a^-}{n^{1/2}},\ \bar{X}_n + \frac{S_{n-1}a^+}{n^{1/2}}\right] \ni m\right) = 1 - \alpha.$$

Cette probabilité se réécrit

$$\mathbb{P}_{m,\sigma}\left(\left[\bar{X}_n - \frac{S_{n-1}a^-}{n^{1/2}},\ \bar{X}_n + \frac{S_{n-1}a^+}{n^{1/2}}\right] \ni m\right) = \mathbb{P}_{m,\sigma}\left(-a^- < \frac{n^{1/2}(m-\bar{X}_n)}{S_{n-1}} < a^+\right)$$

qui est égale à $H_{n-1}(a^+) - H_{n-1}(-a^-)$. ■

En utilisant la symétrie de la loi de Student, les intervalles de confiance bilatéraux symétriques sont donc de la forme

$$\left[\bar{X}_n - \frac{H_{n-1}^{-1}(1-\alpha/2)S_{n-1}}{n^{1/2}}, \bar{X}_n + \frac{H_{n-1}^{-1}(1-\alpha/2)S_{n-1}}{n^{1/2}}\right].$$

5.7.3 Ellipsoïde de confiance pour les estimateurs asymptotiquement normaux

En ouverture, on considère dans cette section une suite d'estimateurs $\hat{\theta}_n$ en dimension $d \in \mathbb{N}^*$ et asymptotiquement normaux :

$$n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{\mathrm{L}} X$$

où X est un vecteur gaussien centré, et l'on supposera sa matrice de covariance $\mathbb{V}\mathrm{ar}(X)$ connue. Comme dans la section précédente, le but est de se ramener à une loi connue, ce qui est rendu possible grâce au résultat de convergence suivant.

Proposition 5.7.5. *Si $\mathbb{V}\mathrm{ar}(X)$ est définie positive, alors la suite de variables aléatoires à valeurs réelles*

$$\left(n(\hat{\theta}_n - \theta)^T \mathbb{V}\mathrm{ar}(X)^{-1}(\hat{\theta}_n - \theta), n \in \mathbb{N}^*\right)$$

converge en loi vers la loi du χ^2 à n degrés de liberté.

Démonstration. Puisque $n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{\mathrm{L}} X$ il s'ensuit que

$$n(\hat{\theta}_n - \theta)^T \mathbb{V}\mathrm{ar}(X)^{-1}(\hat{\theta}_n - \theta) \xrightarrow{\mathrm{L}} X^T \mathbb{V}\mathrm{ar}(X)^{-1} X.$$

et le résultat découle donc de la Proposition 4.6.1. ■

Proposition 5.7.6. *Soit G_n la fonction de répartition de la loi du χ^2 à n degrés de liberté : l'ellipsoïde aléatoire*

$$\Lambda_{n,\alpha} = \left\{ x \in \mathbb{R}^d : n(\hat{\theta}_n - x)^T \mathbb{V}\mathrm{ar}(X)^{-1}(\hat{\theta}_n - x) \leq G_n^{-1}(1-\alpha) \right\}$$

est une région de confiance au niveau asymptotique $1-\alpha$.

Démonstration. Il faut montrer que $\mathbb{P}_\theta(\Lambda_{n,\alpha} \ni \theta) \to 1-\alpha$. Par définition,

$$\mathbb{P}_\theta\left(\Lambda_{n,\alpha} \ni \theta\right) = \mathbb{P}_\theta\left(n(\hat{\theta}_n - \theta)^T \mathbb{V}\mathrm{ar}(X)^{-1}(\hat{\theta}_n - \theta) \leq G_n^{-1}(1-\alpha) \right).$$

La proposition précédente implique donc que

$$\mathbb{P}_\theta\left(\Lambda_{n,\alpha} \ni \theta\right) \underset{n\to\infty}{\longrightarrow} G_n\left(G_n^{-1}(1-\alpha)\right) = 1-\alpha,$$

ce qui montre le résultat attendu. ■

5.8 Fiche de synthèse

Premières définitions

On suppose que $\mathbb{P} \in \{\mathbb{P}_\theta \ ; \ \theta \in \Theta\}$ et qu'il existe un unique $\theta^* \in \Theta$ tel que $\mathbb{P} = \mathbb{P}_{\theta^*}$. Par ailleurs, ou bien chaque $\mathbb{P}_\theta$ est absolument continu de densité f_θ, ou bien chaque $\mathbb{P}_\theta$ est discret de loi p_θ. Dans tous les cas on notera $p(\,\cdot\,;\theta)$ la loi, discrète ou continue, de $\mathbb{P}_\theta$.

- *Échantillon de taille n* : $(X_1, \dots, X_n)$ où les X_k sont indépendantes de même loi $\mathbb{P}_\theta$.
- *Estimateur de θ* : variable aléatoire $\hat{\theta}_n$ de la forme $f(X_1, \cdots, X_n)$ à valeurs dans Θ.
- Estimateur *sans biais* si $\mathbb{E}_\theta(\hat{\theta}_n) = \theta$; *asymptotiquement sans biais* si $\lim\limits_{n\to+\infty} \mathbb{E}_\theta(\hat{\theta}_n) = \theta$.
- Estimateur p.s. convergent si $\hat{\theta}_n \overset{p.s.}{\to} \theta$ quand $n \to +\infty$.
- *Vraisemblance* : $\mathcal{L}(\theta; \mathbf{x}) = p_n(\mathbf{x}; \theta)$: on met l'accent sur le fait que les observations $\mathbf{x}$ sont fixées mais le paramètre θ varie.

Quelques estimateurs

- *Estimateur de la moyenne (moyenne empirique) :* $\boxed{\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k}$; $\mathbb{E}_\theta(\bar{X}_n) = \mathbb{E}_\theta(X)$, $\mathbb{V}\mathrm{ar}_\theta(\bar{X}_n) = \frac{\mathbb{V}\mathrm{ar}_\theta(X)}{n}$.
- *Estimateur du maximum de vraisemblance :* $\boxed{\hat{\theta}_n = \arg\max\limits_{\theta\in\Theta} \mathcal{L}(\theta; X_n)}$

où, si $\vec{x} = (x_1, \cdots, x_n)$, $\mathcal{L}(\theta; \vec{x}) = \prod\limits_{k=1}^{n} p(x_k; \theta)$ avec

$$p(x_k;\theta) = \begin{cases} \mathbb{P}_\theta(X = x_k) & \text{si } X \text{ est discrète} \\ f_\theta(x_k) & \text{si } X \text{ est absolument continue.} \end{cases}$$

Remarque : $\theta \mapsto \mathcal{L}(\theta, \vec{x})$ et $\theta \mapsto \ln \mathcal{L}(\theta, \vec{x})$ ayant les mêmes variations, la deuxième expression se prête souvent mieux à la dérivation.

Rappel de Conditionnement :

- *Espérance totale* : Si Y admet une espérance, $\mathbb{E}(Y|X)$ aussi et alors $\boxed{\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y)}$.
- *Propriété* : $\mathbb{E}(\psi(X)Y|X) = \psi(X)\mathbb{E}(Y|X)$ pour toute fonction ψ.
- *Vecteur gaussien* : Si $Z = (Z_1, Z_2)$ est un vecteur gaussien tel que Z_1 est absolument continu, alors

$$\boxed{\mathbb{E}(Z_2 \mid Z_1) = \mathbb{E}(Z_2) + \mathbb{C}\mathrm{ov}(Z_2, Z_1)\mathbb{V}\mathrm{ar}(Z_1)^{-1}(Z_1 - \mathbb{E}(Z_1))}.$$

Définitions et propriétés supplémentaires :

- *Information de Fisher* : $\boxed{I(\theta) = \mathbb{V}\mathrm{ar}_\theta\left(\partial_\theta \ln p(X_1;\theta)\right)}$.

Propriété : Si le modèle est régulier (voir p88), on a aussi $\boxed{I(\theta) = -\mathbb{E}_\theta\left(\partial_\theta^2 \ln p(X_1;\theta)\right)}$.
Théorème 5.5.1 : Si le modèle est régulier, l'estimateur du maximum de vraisemblance est convergent en loi.
Borne de Fréchet–Darmois–Cramer–Rao (Théorème 5.6.1) : Si le modèle est régulier et $I(\theta) > 0$ alors $\boxed{\mathbb{V}\mathrm{ar}_\theta(\hat{\theta}_n) \geq \frac{1}{nI(\theta)}}$.

- Estimateur *efficace* si $\mathbb{V}\mathrm{ar}_\theta(\hat{\theta}_n) = \frac{1}{nI(\theta)}$ (*asymptotiquement* si $\lim\limits_{n\to+\infty} n\mathbb{V}\mathrm{ar}_\theta(\hat{\theta}_n) = \frac{1}{I(\theta)}$).

Théorème 5.6.2 : Si le modèle est identifiable et régulier, alors l'estimateur du maximum de vraisemblance est asymptotique normal, i.e., pour tout θ, sous $\mathbb{P}_\theta$, $\boxed{\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathrm{L}} X \text{ de loi normale } \mathcal{N}(0, 1/I(\theta))}$.

• Pour $\alpha \in [0,1]$, Λ_α est une *région de confiance de niveau* $1-\alpha$ si $\boxed{\mathbb{P}_\theta(\theta \in \Lambda_\alpha) \geq 1-\alpha}$ pour tout θ.
Exemple : Pour $\mathcal{N}(\theta, \sigma^2)$, $\left[\overline{X}_n - \sigma F^{-1}(1-\alpha/2)/\sqrt{n}, \overline{X}_n + \sigma F^{-1}(1-\alpha/2)/\sqrt{n}\right]$, où F est la fonction de répartition de la loi normale $\mathcal{N}(0,1)$, est l'intervalle de confiance bilatéral symétrique de niveau $1-\alpha$.

5.9 Exercices

Les exercices précédés d'une flèche ↪ sont des exercices d'application directs du cours.

↪ **Exercice 5.1** *(Estimateurs de la moyenne)*
Soit $X_1, \dots, X_n$ n variables aléatoires indépendantes de même loi et de carré intégrable. On s'intéresse dans cet exercice aux estimateurs $\hat{\theta}$ de la moyenne $\theta = \mathbb{E}(X_1)$ qui sont de la forme $\hat{\theta}_n = \sum_{k=1}^n a_k X_k$.

1. Sous quelles conditions sur les coefficients a_k l'estimateur $\hat{\theta}_n$ est-il un estimateur sans biais de la moyenne $\mathbb{E}(X_1)$?

2. En admettant que $(1/n, \dots, 1/n)$ minimise la somme $\sum_{k=1}^n a_k^2$ sous la contrainte $\sum_{k=1}^n a_k = 1$, montrez que la moyenne empirique est un estimateur de la moyenne de variance minimale parmi les estimateurs sans biais de la forme $\sum_{k=1}^n a_k X_k$.

3. (Facultatif) Montrez que $(1/n, \dots, 1/n)$ minimise la somme $\sum_{k=1}^n a_k^2$ sous la contrainte $\sum_{k=1}^n a_k = 1$.

↪ **Exercice 5.2** *(Estimateur du maximum de vraisemblance)*
Calculez l'estimateur du maximum de vraisemblance pour le modèle paramétrique $\{\mathbb{P}_\theta : \theta \in \Theta\}$ dans les cas suivants.

1. $\mathbb{P}_\theta$ est la loi beta $\beta(1, 1/\theta)$: $f_\theta(x) = \frac{1}{\theta}(1-x)^{1/\theta - 1}\mathbb{1}\{0 \le x \le 1\}$.

2. $\mathbb{P}_\theta$ est la loi de Poisson de paramètre θ : $p_\theta(x) = e^{-\theta}\frac{\theta^x}{x!}$ pour $x \in \mathbb{N}$.

3. $\mathbb{P}_\theta$ est la normale de moyenne m fixée et connue et de variance θ^2 : $f_\theta(x) = (2\pi\theta^2)^{-1/2}\exp(-\frac{(x-m)^2}{2\theta^2})$.

4. $\mathbb{P}_\theta$ est la loi uniforme sur $[0, \theta]$: $f_\theta(x) = \frac{1}{\theta}\mathbb{1}\{0 \le x \le \theta\}$.

↪ **Exercice 5.3** *(Estimateurs de la loi uniforme)*
La durée d'un feu rouge est égale à θ^*, paramètre inconnu strictement positif. On observe un échantillon $T_1, T_2, \dots, T_n$ de taille n où T_i désigne la durée d'attente du i-ème automobiliste. On suppose que les T_i sont i.i.d. et suivent une loi uniforme sur $[0, \theta^*]$.

1. Formulez un modèle paramétrique $\{\mathbb{P}_\theta, \theta > 0\}$ visant à estimer θ^*.

2. Calculez $\mathbb{E}_\theta(T_1)$ et $\mathbb{V}\mathrm{ar}_\theta(T_1)$ et déduisez-en que $2T_1$ est un estimateur sans biais de θ. Est-il convergent ?

3. On considère $\bar{T}_n = (1/n)\sum_{k=1}^n T_k$ la moyenne empirique. Calculez $\mathbb{E}_\theta(\bar{T}_n)$ et $\mathbb{V}\mathrm{ar}_\theta(\bar{T}_n)$ et déduisez-en que $\hat{\theta}_n^{(1)} = 2\bar{T}_n$ est un estimateur sans biais de θ. Est-il convergent ?

4. Calculez l'estimateur du maximum de vraisemblance M_n.

5. Calculez la loi de M_n sous $\mathbb{P}_\theta$ et déduisez-en $\mathbb{E}_\theta(M_n)$ puis $\mathbb{V}\mathrm{ar}_\theta(M_n)$. L'estimateur M_n est-il sans biais ? convergent ?

6. Quel estimateur choisiriez-vous pour estimer θ^* ? Justifiez votre réponse.

↪ **Exercice 5.4** *(Estimateur de la loi de Bernoulli)*
On considère une chaîne de production à la sortie de laquelle chaque pièce a une probabilité p^* d'être défectueuse, indépendamment les unes des autres.

1. Pour estimer p^*, on prélève un échantillon de n pièces et on définit X_i la variable aléatoire qui vaut 1 si la i-ième pièce de l'échantillon est défectueuse et 0 sinon. Formulez un modèle paramétrique adéquat pour estimer p^*, calculez l'estimateur du maximum de vraisemblance $\hat{\theta}$ associé ainsi que sa variance.
Indication. On pourra utiliser la formule $\mathbb{P}(X = x) = p^x(1-p)^{1-x}$ valable pour une variable aléatoire de Bernoulli de paramètre p.

On répète cette procédure deux fois mais avec des tailles d'échantillon différentes : le premier échantillon a une taille n_1 et le deuxième une taille n_2. On note $\hat{\theta}^{(1)}$ l'estimateur du maximum de vraisemblance du premier échantillon et $\hat{\theta}^{(2)}$ celui du second.

2. $(\hat{\theta}^{(1)} + \hat{\theta}^{(2)})/2$ est-il un estimateur sans biais de p ? Dans la classe des estimateurs de la forme $a_1\hat{\theta}^{(1)} + a_2\hat{\theta}^{(2)}$, trouvez par le calcul celui sans biais et de variance minimale.

3. Retrouvez directement ce résultat à l'aide de l'exercice 1 en considérant $X_k^{(i)} = 1$ si la k-ième pièce du i-ième échantillon est défectueuse et 0 sinon.

↪ **Exercice 5.5** *(Loi de Bernoulli)*
Soit $\mathbb{P}_\theta$ pour $\theta \in [0,1]$ la loi de Bernoulli de paramètre θ. On rappelle (cf. exercice 5.4) que l'estimateur du maximum de vraisemblance associé est donné par la moyenne empirique $\hat{\theta}_n = (1/n)\sum_{k=1}^n X_k$.

1. Calculez le biais et l'erreur quadratique moyenne de $\hat{\theta}_n$. Cet estimateur est-il convergent, efficace ?

2. Quelle est la limite de $\sqrt{n}(\hat{\theta}_n - \theta)$? Comparez ce résultat au théorème 5.6.2 du cours.

3. Pour $m \in [0,1]$ montrez que la fonction

$$p \in (0,1) \mapsto \frac{m-p}{\sqrt{p(1-p)}}$$

est décroissante. Déduisez-en que pour tout $a_- < a_+$ et $m \in [0,1]$, l'ensemble

$$I(m, a_-, a_+) = \left\{ p \in (0,1) : a_- \leq \frac{m-p}{\sqrt{p(1-p)}} \leq a_+ \right\}$$

est un intervalle, puis utilisez la question précédente pour montrer que $I(\hat{\theta}_n, a_+/\sqrt{n}, a_-/\sqrt{n})$ est un intervalle de confiance de niveau asymptotique $F(a_+) - F(a_-)$.

4. *Question bonus.* Pour $0 \leq r \leq n$ entiers et $\alpha \in (0,1)$, on considère $p_1 = p_1(r,n,\alpha)$ et $p_2 = p_2(r,n,\alpha)$ tels que

$$\sum_{k=r}^{n} \binom{n}{k} p_1^k (1-p_1)^{n-k} = \frac{\alpha}{2} \quad \text{et} \quad \sum_{k=0}^{r} \binom{n}{k} p_2^k (1-p_2)^{n-k} = \frac{\alpha}{2}.$$

Montrez que $[p_1(n\hat{\theta}_n, n, \alpha), p_2(n\hat{\theta}_n, n, \alpha)]$ est un intervalle de confiance de niveau $1-\alpha$ pour θ. Quel est son inconvénient ?

↪ **Exercice 5.6** *(Loi beta)*
Soit $\mathbb{P}_\theta$ pour $\theta > 0$ la loi beta $\beta(1, 1/\theta)$. On rappelle qu'il s'agit de la loi de densité

$$f_\theta(x) = \frac{1}{\theta}(1-x)^{1/\theta - 1} \mathbb{1}\{x \in [0,1]\}$$

et (cf. exercice 2 de la séance précédente) que l'estimateur du maximum de vraisemblance associé est donné par $\hat{\theta}_n = -(1/n)\sum_{k=1}^n \log(1-X_k)$.

1. Montrez que $-\log(1-X_i)$ sous $\mathbb{P}_\theta$ suit une loi exponentielle : quel est son paramètre ? Déduisez-en sa moyenne et sa variance.

2. Calculez le biais et l'erreur quadratique moyenne de $\hat{\theta}_n$. Cet estimateur est-il convergent, efficace ?

3. Soit G_n la fonction de répartition de la loi Gamma(n,n). Montrez que $\hat{\theta}_n/\theta$ suit une loi Gamma(n,n) et déduisez-en une famille d'intervalles de confiance au niveau $1-\alpha$ pour θ exprimée en fonction de G_n.
Indications. On rappelle que la loi Gamma(n,λ) est la loi de la somme de n variables exponentielles i.i.d. de paramètre λ, et que si E est une variable exponentielle de paramètre λ alors μE est une variable exponentielle de paramètre λ/μ.

4. Montrez que $\sqrt{n}(\hat{\theta}_n/\theta - 1)$ sous $\mathbb{P}_\theta$ converge en loi : quelle est la limite ? Comparez au théorème 5.6.2 du cours et déduisez-en un intervalle de confiance asymptotique.

↪ **Exercice 5.7** *(Loi uniforme)*
Soit $\mathbb{P}_\theta$ pour $\theta > 0$ la loi uniforme sur $[0,\theta]$. On rappelle (cf. exercice 2 de la séance précédente) que l'estimateur du maximum de vraisemblance associé est $\hat{\theta}_n = \max_{k=1,\dots,n} X_k$.

1. Peut-on utiliser la borne de Fréchet–Darmois–Cramer–Rao ? Pourquoi ?

2. Montrez que la loi de $\hat{\theta}_n/\theta$ sous $\mathbb{P}_\theta$ est égale à la loi le $\hat{\theta}_n$ sous $\mathbb{P}_1$, et déduisez-en que $[\hat{\theta}_n/a_+, \hat{\theta}_n/a_-]$ est un intervalle de confiance pour θ au niveau $1-\alpha$ pour tout a_-, a_+ satisfaisant $a_+^n - a_-^n = 1-\alpha$.

3. Montrez que $n(1-\hat{\theta}_n/\theta)$ sous $\mathbb{P}_\theta$ converge en loi vers une loi exponentielle dont on identifiera le paramètre, et utilisez ce résultat pour construire un intervalle de confiance au niveau asymptotique $1-\alpha$.

Problème 5.8 *(Statistique bayésienne)*
On s'intéresse ici à l'estimation d'une variable aléatoire Θ à valeur dans $\mathbb{R}$ à l'aide de mesures bruitées. On suppose plus précisément que la i-ème observation X_i est donnée par

$$X_i = \Theta + W_i$$

où les W_i sont i.i.d. et suivent une loi normale centrée et de variance σ_w^2 supposée connue.

1. Montrez que pour toute fonction $g : \mathbb{R}^n \to \mathbb{R}$ mesurable, on a

$$\mathbb{E}\big[\big(g(\mathbf{X}_n) - \Theta\big)^2 \mid \mathbf{X}_n\big] = \big(g(\mathbf{X}_n) - \mathbb{E}(\Theta \mid \mathbf{X}_n)\big)^2 + \mathbb{E}\big[\big(\mathbb{E}(\Theta \mid \mathbf{X}_n) - \Theta\big)^2 \mid \mathbf{X}_n\big]$$

et déduisez-en que

$$\mathbb{E}\big[\big(\mathbb{E}(\Theta \mid \mathbf{X}_n) - \Theta\big)^2\big] \leq \mathbb{E}\big[\big(g(\mathbf{X}_n) - \Theta\big)^2\big].$$

Ce résultat dépend-il de la loi des W_i ? Justifiez la terminologie *estimateur des moindres carrés* pour $\mathbb{E}(\Theta \mid \mathbf{X}_n)$.

Par la suite, on appelle **erreur quadratique moyenne** $\mathrm{EQM}(\hat{\Theta})$ associée à un estimateur $\hat{\Theta}$ la distance L_2 au carré entre $\hat{\Theta}$ et Θ :

$$\mathrm{EQM}(\hat{\Theta}) = \mathbb{E}\big[\big(\hat{\Theta} - \Theta\big)^2\big].$$

Première partie. Dans la première partie du problème, on suppose que $\Theta = \theta_0 \in \mathbb{R}$ est une variable déterministe (i.e., une constante).

2. Que vaut $\mathbb{E}(\Theta \mid \mathbf{X}_n)$ dans ce cas ? $\mathbb{E}(\Theta \mid \mathbf{X}_n)$ est-il vraiment un estimateur ?

3. Montrez que l'estimateur du maximum de vraisemblance est égal à la moyenne empirique

$$\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

$\bar{X}_n$ est-il convergent ? quelle est sa limite ?

4. Montrez que son erreur quadratique moyenne est donné par

$$\mathrm{EQM}(\bar{\mathrm{X}}_\mathrm{n}) = \frac{\sigma_\mathrm{w}^2}{\mathrm{n}}.$$

Cette erreur quadratique moyenne est-elle plus ou moins élevée que l'erreur quadratique moyenne de l'espérance conditionnelle ?

Deuxième partie. Dans la deuxième partie du problème, on suppose que Θ suit une loi normale de moyenne $\mathbb{E}(\Theta)$ et de variance $\mathbb{V}\mathrm{ar}(\Theta)$ et est indépendante des W_i.

5. Quel est le comportement asymptotique de la moyenne empirique $\bar{X}_n$?

6. Montrez que $(\Theta, \mathbf{X}_n)$ est un vecteur gaussien dont la matrice de variance-covariance est donnée par

$$\mathbb{V}\mathrm{ar}((\Theta, \mathbf{X}_n)) = \begin{pmatrix} \mathbb{V}\mathrm{ar}(\Theta) & \mathbb{V}\mathrm{ar}(\Theta) & \mathbb{V}\mathrm{ar}(\Theta) & \cdots & \mathbb{V}\mathrm{ar}(\Theta) \\ \mathbb{V}\mathrm{ar}(\Theta) & \mathbb{V}\mathrm{ar}(\Theta) + \sigma_w^2 & \mathbb{V}\mathrm{ar}(\Theta) & \cdots & \mathbb{V}\mathrm{ar}(\Theta) \\ \mathbb{V}\mathrm{ar}(\Theta) & \mathbb{V}\mathrm{ar}(\Theta) & \mathbb{V}\mathrm{ar}(\Theta) + \sigma_w^2 & \cdots & \mathbb{V}\mathrm{ar}(\Theta) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbb{V}\mathrm{ar}(\Theta) & \mathbb{V}\mathrm{ar}(\Theta) & \mathbb{V}\mathrm{ar}(\Theta) & \cdots & \mathbb{V}\mathrm{ar}(\Theta) + \sigma_w^2 \end{pmatrix}.$$

7. Déduisez à l'aide du Théorème 4.5.1 du polycopié que

$$\mathbb{E}(\Theta \mid \mathbf{X}_n) = \alpha_n \bar{X}_n + (1-\alpha_n)\mathbb{E}(\Theta) \text{ avec } \alpha_n = \frac{n\mathbb{V}\mathrm{ar}(\Theta)}{n\mathbb{V}\mathrm{ar}(\Theta) + \sigma_w^2}.$$

Indication : Si H est la matrice de taille n dont toutes les entrées valent 1 et I la matrice identité de taille n aussi, alors pour tout $a, b > 0$ on a on a $(aH + bI)^{-1} = a'H + b'I$ avec $a' = -a/(b^2 + nab)$ et $b' = 1/b$.

8. Sous quelles conditions peut-on se servir de $\mathbb{E}(\Theta \mid \mathbf{X}_n)$ comme estimateur ? Interprétez les valeurs de $\mathbb{E}(\Theta \mid \mathbf{X}_n)$ pour n petit et grand.

9. Montrez que l'erreur quadratique moyenne associée à $\mathbb{E}(\Theta \mid \mathbf{X}_n)$ est donnée par

$$\mathrm{EQM}(\mathbb{E}(\Theta \mid \mathbf{X}_\mathrm{n})) = \frac{\alpha_\mathrm{n}^2 \sigma_\mathrm{w}^2}{\mathrm{n}} + (1 - \alpha_\mathrm{n})^2 \mathbb{V}\mathrm{ar}(\Theta)$$

et vérifiez que l'on a bien $\mathrm{EQM}(\mathbb{E}(\Theta \mid \mathbf{X}_\mathrm{n})) \leq \mathrm{EQM}(\bar{\mathrm{X}}_\mathrm{n})$.

Troisième partie. On suppose toujours que Θ suit une loi normale et est indépendante des W_i mais on ne connaît ni sa moyenne ni sa variance. On fixe $\mu \in \mathbb{R}$ et une suite $a_n \in [0, 1]$ et l'on considère

$$\hat{\Theta}_n = a_n \bar{X}_n + (1 - a_n)\mu.$$

10. $\hat{\Theta}_n$ est-il un estimateur de Θ ? à quelle condition est-il convergent ?

11. Montrez que l'erreur quadratique moyenne associée à $\hat{\Theta}_n$ est donnée par

$$\mathrm{EQM}(\hat{\Theta}_\mathrm{n}) = \frac{\mathrm{a}_\mathrm{n}^2 \sigma_\mathrm{w}^2}{\mathrm{n}} + (1 - \mathrm{a}_\mathrm{n})^2 \left[(\mathbb{E}(\Theta) - \mu)^2 + \mathbb{V}\mathrm{ar}(\Theta) \right].$$

12. Comparez $\mathrm{EQM}(\bar{\mathrm{X}}_\mathrm{n})$, $\mathrm{EQM}(\mathbb{E}(\Theta \mid \mathbf{X}_\mathrm{n}))$ et $\mathrm{EQM}(\hat{\Theta}_\mathrm{n})$: sous quelles conditions sur μ et a a-t-on $\mathrm{EQM}(\hat{\Theta}_\mathrm{n}) \leq \mathrm{EQM}(\bar{\mathrm{X}}_\mathrm{n})$?

Quatrième partie. On suppose maintenant que n varie dans le temps et que les observations $X_1, X_2, \ldots$, sont acquises une par une séquentiellement.

13. Montrez que

$$\mathbb{E}(\Theta \mid \mathbf{X}_{n+1}) = \left(1 - \frac{\mathbb{V}\mathrm{ar}(\Theta)}{\sigma_w^2 + (n+1)\mathbb{V}\mathrm{ar}(\Theta)}\right) \mathbb{E}(\Theta \mid \mathbf{X}_n) + \frac{\mathbb{V}\mathrm{ar}(\Theta)}{\sigma_w^2 + (n+1)\mathbb{V}\mathrm{ar}(\Theta)} X_{n+1}.$$

Quel est l'intérêt ce cette formule ?

Problème 5.9

On considère 5 groupes de femmes âgées respectivement de 35, 45, 55, 65 et 75 ans. Dans chaque groupe, on a mesuré la tension artérielle en mm de mercure de chaque et on a calculé la valeur moyenne pour chaque groupe. On définit donc les variables :

T : tension moyenne en mm Hg	114	124	143	158	166
x : âge du groupe considéré	35	45	55	55	75

TABLEAU 5.1 – Tableau donnant la tension moyenne en fonction de l'âge du groupe.

Afin de visualiser ces données, on fait une représentation cartésienne comme sur la Figure 5.1.
Sur le graphique, on constate que la tension artérielle augmente avec l'âge, mais surtout, que cette augmentation semble linéaire puisque les points du graphique sont presque alignés. On considère donc le modèle suivant pour expliquer cette relation de dépendance :

$$T_i = \alpha^* + \beta^* x_i + \varepsilon_i$$

où les ε_i représentent un bruit par rapport au modèle déterministe strictement linéaire. D'un point de vue technique, on supposera que les ε_i sont décorrélées, centrées et de même variance σ^2, i.e., pour tout $i \neq j$,

$$\mathbb{E}(\varepsilon_i) = 0, \ \mathbb{V}\mathrm{ar}(\varepsilon_i) = \sigma^2 \ \text{ et } \ \mathbb{C}\mathrm{ov}(\varepsilon_i, \varepsilon_j) = 0.$$

1. Est-on dans le cadre d'application du cours ? Par exemple, le Théorème 5.6.1 s'applique-t-il ?

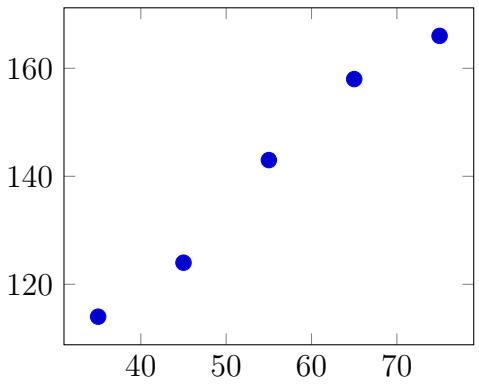


FIGURE 5.1 – Représentation graphique de la relation entre âge et tension artérielle.

On considère l'estimateur des moindres carrés $\hat{\theta}_n = (\hat{\alpha}_n, \hat{\beta}_n)$ donné par

$$\hat{\theta}_n \in \arg\min \left\{ \sum_{i=1}^{n} (T_i - \alpha - \beta x_i)^2 : \alpha, \beta \in \mathbb{R} \right\}.$$

Pour $\mathbf{x}$ et $\mathbf{t}$ deux vecteurs de longueur n, on définit les moyenne, corrélation et variance empirique par

$$m(\mathbf{x}) = \frac{1}{n}\sum_{k=1}^{n} x_k, \;\; C(\mathbf{x},\mathbf{t}) = \frac{1}{n-1}\sum_{k=1}^{n} (x_k - m(\mathbf{x}))\,(t_k - m(\mathbf{t})) \quad \text{et} \quad V(\mathbf{x}) = C(\mathbf{x},\mathbf{x})$$

et par la suite on note $\mathbf{T}_n = (T_1, \ldots, T_n)$ et $\mathbf{x}_n = (x_1, \ldots, x_n)$.

2. Montrez que $\hat{\beta}_n = C(\mathbf{x}_n, \mathbf{T}_n)/V(\mathbf{x}_n)$ et que $\hat{\alpha}_n = m(\mathbf{T}_n) - \hat{\beta}_n m(\mathbf{x}_n)$.

3. Montrez que $\hat{\theta}_n = (\hat{\alpha}_n, \hat{\beta}_n)$ est sans biais. Par la suite on admettra que sa matrice de variance-covariance est donnée par

$$\mathbb{V}\mathrm{ar}_\theta(\hat{\theta}_n) = \frac{\sigma^2}{(n-1)V(\mathbf{x}_n)} \begin{pmatrix} m(\mathbf{x}_n^2) & -m(\mathbf{x}_n) \\ -m(\mathbf{x}_n) & 1 \end{pmatrix}$$

et on supposera en plus des hypothèses précédentes que les ε_i suivent une loi normale.

4. Quelle est la loi de $\mathbf{T}_n$? Donnez la formule de sa densité $p_n(\,\cdot\,;\theta)$ sous $\mathbb{P}_\theta$ et déduisez-en que $\hat{\theta}_n$ coïncide avec l'estimateur du maximum de vraisemblance.

5. Quelle est la loi de $\hat{\theta}_n$? Déduisez-en des intervalles de confiance pour α^* et β^*.

6. Peut-on utiliser ces intervalles lorsque σ est inconnu ? Montrez que dans ce cas, l'estimateur du maximum de vraisemblance de σ est donné par

$$\hat{\sigma}_n = \sqrt{\frac{1}{n}\sum_{k=1}^{n} \left(T_k - \hat{\alpha}_n - \hat{\beta}_n x_k\right)^2}.$$

On admet que le Théorème 5.6.1 se généralise de la manière suivante : pour tout estimateur $\hat{\theta}$, on a l'inégalité matricielle $\mathbb{V}\mathrm{ar}_\theta(\hat{\theta}) \geq [\mathbb{E}_\theta(M_n(\theta))]^{-1}$ où $M_n(\theta)$, supposée inversible, est la matrice aléatoire définie par (on revient au cas où σ est supposé connu)

$$M_n(\theta) = -\nabla_\theta^2 \log p_n(\mathbf{T}_n;\theta) = -\begin{pmatrix} \dfrac{\partial^2 \log p_n}{\partial \alpha^2} & \dfrac{\partial^2 \log p_n}{\partial \alpha \partial \beta} \\ \dfrac{\partial^2 \log p_n}{\partial \alpha \partial \beta} & \dfrac{\partial^2 \log p_n}{\partial \beta^2} \end{pmatrix} (\mathbf{T}_n;\theta).$$

On dira qu'un estimateur est efficace si sa variance est égale à $[\mathbb{E}_\theta(M_n(\theta))]^{-1}$.

7. Calculez $M_n(\theta)$ puis $[\mathbb{E}_\theta(M_n(\theta))]^{-1}$: l'estimateur $\hat{\theta}_n$ est-il efficace ?

8. On suppose maintenant que σ est inconnu et on utilise les estimateurs $\hat{\alpha}_n$, $\hat{\beta}_n$ et $\hat{\sigma}_n$ pour faire de la prédiction. Etant donnée une valeur $z \notin \{x_1, \ldots, x_n\}$, on considère $\hat{Y}_n(z) = \hat{\alpha}_n + \hat{\beta}_n z$ la prédiction de la tension artérielle pour un groupe d'âge z. On admettra que l'intervalle de confiance symétrique au niveau $1-\alpha$ pour $\hat{Y}_n(z)$ est donné par

$$\left[\hat{\alpha}_n + \hat{\beta}_n z - t_{n,\alpha}\hat{\sigma}_n^*, \hat{\alpha}_n + \hat{\beta}_n z + t_{n,\alpha}\hat{\sigma}_n^*\right]$$

où

$$\hat{\sigma}_n^* = \frac{n^{1/2}\hat{\sigma}_n}{(n-2)^{1/2}} \sqrt{1 + \frac{1}{n} + \frac{(z - m(\mathbf{x}_n))^2}{(n-1)V(\mathbf{x}_n)}}$$

et $t_{n,\alpha}$ est le quantile d'ordre $1-\alpha/2$ pour la loi de Student à $n-2$ degrés de liberté. Quelles sont les valeurs z le mieux prédites par ce modèle ? Interprétez ce résultat.

Chapitre 6

Tests d'hypothèses

Pour expliquer les concepts importants derrière les tests d'hypothèse, on commence par décortiquer dans la Section 6.1 un exemple très simple que l'on formalise dans le reste du chapitre.

6.1 Une histoire de biais

Pour motiver le concept d'estimation paramétrique, nous avons considéré l'exemple de l'estimation du biais d'une pièce de monnaie. Continuous cet exemple et imaginons qu'en fait, j'ai acheté cette pièce pour faire des tours de magie et le vendeur m'a assuré qu'elle était biaisée, avec une probabilité $p = 0{,}54$ d'obtenir pile. Au-delà d'estimer le biais de la pièce, je souhaite plus précisément **tester l'hypothèse** selon laquelle la pièce a biais 0,54 : le but de ce chapitre est de répondre à une telle question.

6.1.1 Première approche : une histoire de seuil

Grâce au chapitre précédent, on sait dorénavant que si on considère une série de lancers et que l'on définit $X_k = 1$ si l'on obtient pile au k-ième lancer et 0 sinon, alors la moyenne empirique $\bar{X}_n$ est un bon estimateur du biais (il est sans biais, convergent et efficace). Une idée naturelle est donc de choisir la règle de décision suivante :

Règle de décision informelle : On décide que la pièce n'a pas un biais 0,54 si l'écart $|\bar{X}_n - 0{,}54|$ entre estimation $\bar{X}_n$ et valeur testée 0,54 est grand.

Immédiatement vient alors la question de ce que l'on entend par "grand" : doit-on décider que le biais ne vaut pas 0,54 si $|\bar{X}_n - 0{,}54|$ vaut $1/20$? $1/10$? Comment choisir ce seuil ?

Pour répondre à cette question, il est impératif de comprendre que cette procédure possède un risque inhérent : même si le biais de la pièce valait 0,54, il serait toujours possible d'obtenir $\bar{X}_n = 0$: cela arriverait avec la probabilité $0{,}46^n$ qui, bien que faible, est strictement positive. Ainsi,

Le seuil auquel on décidera que la pièce est biaisée est déterminé par le risque de se tromper que l'on est prêts à prendre.

En d'autres termes, le seuil est déterminé par une assertion du genre :

Je veux fixer mon seuil afin de ne me tromper qu'avec au plus 5% *de chance.*

Ici, *se tromper* veut dire que, **bien que la pièce a le biais annoncé**, on rejette cette hypothèse car la distance $|\bar{X}_n - 0{,}54|$ est trop grande, supérieure au seuil κ fixé. Ainsi, pour calculer la probabilité de se tromper il faut **faire les calculs sous** $\mathbb{P}_{0,54}$ – $\mathbb{P}_{0,54}$ est la mesure de probabilité sous laquelle le biais de la pièce vaut 0,54, qui correspond à supposer que le biais de la pièce vaut 0,54 – et l'on trouve donc que la probabilité de se tromper vaut

$$\mathbb{P}_{0,54}\left(|\bar{X}_n - 0{,}54| \geq \kappa\right).$$

Si l'on veut fixer le seuil κ afin de ne se tromper qu'avec 5% de chance, cela revient donc à résoudre l'équation

$$\mathbb{P}_{0,54}\left(|\bar{X}_n - 0{,}54| \geq \kappa\right) = 5\%. \tag{6.1}$$

L'équation (6.1) n'est pas évidente à résoudre de manière exacte : par contre, une approximation est donnée par le théorème central limite qui nous assure que $n^{1/2}(\bar{X}_n - 0{,}54)/\sigma$ sous $\mathbb{P}_{0,54}$ et avec $\sigma^2 = 0,54 \times 0{,}46$ converge en distribution vers une loi normale standard. Ainsi, pour n raisonnablement grand on a l'approximation

$$\mathbb{P}_{0,54}\left(|\bar{X}_n - 0{,}54| \geq \kappa\right) \approx \mathbb{P}\left(|X| \geq \frac{n^{1/2}\kappa}{\sigma}\right)$$

où X suit une loi normale standard. Si F est la fonction de répartition de X, κ est donc (approximativement) déterminé par la relation

$$2\left(1 - F\left(\frac{n^{1/2}\kappa}{\sigma}\right)\right) = 5\%$$

soit

$$\kappa = \frac{\sigma}{n^{1/2}} F^{-1}(0{,}975).$$

Pour ce choix du seuil, on a donc conçu un **test statistique** qui, pour $n = 50$ lancers, peut être résumé de la façon suivante :

Règle de décision #1 : On rejettera l'hypothèse que la pièce a un biais de 0,54 si

$$|\bar{X}_{50} - 0{,}54| > \frac{\sqrt{0{,}54 \times 0{,}46} \times F^{-1}(0{,}975)}{\sqrt{50}} \approx 0{,}138.$$

En outre, on a conçu le test de telle sorte que, si la pièce a effectivement un biais de 0,54, alors la probabilité de se tromper (i.e., de décider que la pièce n'a pas un biais de 0,54) est de 5%.

6.1.2 Deuxième approche : intervalles de confiance

La procédure ci-dessus fait fortement penser aux raisonnements que l'on a effectués lors de la construction d'intervalles de confiance au chapitre précédent. Et effectivement, le test statistique proposé ci-dessus peut se reformuler en termes d'intervalle de confiance de la manière suivante.

Considérons I l'intervalle de confiance bilatéral symétrique au niveau $\alpha \in (0,1)$ de la probabilité de succès : sous l'approximation normale effectuée ci-dessus, on sait grâce à la Proposition 5.7.1 que I_α est donné par

$$I_\alpha = \left[\bar{X}_n - \frac{\sigma F^{-1}(1-\alpha/2)}{n^{1/2}}, \bar{X}_n + \frac{\sigma F^{-1}(1-\alpha/2)}{n^{1/2}}\right].$$

Il est alors naturel de considérer le test statistique suivant :

Règle de décision #2 : On décide que la pièce n'a pas un biais 0,54 si $0{,}54 \notin I_{5\%}$.
La probabilité de se tromper est alors

$$\mathbb{P}_{0,54}\left(0{,}54 \notin I_{5\%}\right)$$

qui par construction de l'intervalle de confiance vaut 5%. En outre, on observe que les deux règles de décision proposées sont les mêmes ! En d'autres termes, on peut se servir des intervalles de confiance pour construire des tests statistiques.

6.1.3 Troisième approche : niveau de signification

Nous présentons maintenant une troisième interprétation du même test statistique qui s'appuie sur la notion de niveau de signification. L'idée est de concevoir un test selon le critère suivant. Etant donnée la réalisation de $\bar{X}_n$, on va calculer la **probabilité d'obtenir une valeur au moins aussi extrême que** $\bar{X}_n$: si cette probabilité est faible, on rejettera le test.

Pour rendre cette idée plus concrète, imaginons qu'on observe $\bar{X}_n = 0{,}7$: cela paraît un résultat très large, au moins comparé au 0,54 attendu, et l'idée est de calculer la probabilité $\mathbb{P}_{0,54}(\bar{X}_n \geq 0{,}7)$ d'obtenir un résultat au moins aussi extrême. Si inversement on avait observé $\bar{X}_n = 0{,}2$, on se serait intéressés à la probabilité $\mathbb{P}_{0,54}(\bar{X}_n \leq 0{,}2)$, et dans les deux cas on aurait rejeté l'hypothèse que le biais vaut 0,54 si la probabilité calculée était faible. En effet, cela signifierait que l'hypothèse selon laquelle le biais vaut 0,54 n'explique pas bien les données, et il est donc naturel de la rejeter.

On montre maintenant que ce test est, à un facteur multiplicatif près, équivalent au test précédent. Par définition, on a proposé la règle de décision suivante :

Règle de décision #3 : On décide que la pièce n'a pas un biais 0,54 si $P \leq \alpha$ où

$$P = \mathbb{1}\left\{\bar{X}_n < 0{,}54\right\}\mathbb{P}_{0,54}\left(\bar{X}'_n \leq \bar{X}_n \mid \bar{X}_n\right) + \mathbb{1}\left\{\bar{X}_n > 0{,}54\right\}\mathbb{P}_{0,54}\left(\bar{X}'_n \geq \bar{X}_n \mid \bar{X}_n\right)$$

avec $\bar{X}'_n$ de même loi que $\bar{X}_n$ et indépendante de $\bar{X}_n$.

On laisse effectivement le lecteur se convaincre que la probabilité (aléatoire) P est exactement la probabilité d'obtenir une valeur au moins aussi extrême que $\bar{X}_n$. En outre, on a

$$\mathbb{P}_{0,54}\left(\bar{X}'_n \leq \bar{X}_n \mid \bar{X}_n\right) = \mathbb{P}_{0,54}\left(Z'_n \leq Z_n \mid \bar{X}_n\right)$$

où l'on a défini

$$Z_n = \frac{n^{1/2}(\bar{X}_n - 0{,}54)}{\sigma} \quad \text{et} \quad Z'_n = \frac{n^{1/2}(\bar{X}'_n - 0{,}54)}{\sigma},$$

et donc sous l'approximation de normalité du théorème central limite, on obtient

$$\mathbb{P}_{0,54}\left(\bar{X}'_n \leq \bar{X}_n \mid \bar{X}_n\right) \approx F(Z_n).$$

Sous cette approximation, on a donc

$$P = \mathbb{1}\left\{\bar{X}_n < 0{,}54\right\}F(Z_n) + \mathbb{1}\left\{\bar{X}_n > 0{,}54\right\}F(-Z_n)$$

et l'on vérifie que

$$P \leq \alpha \iff |\bar{X}_n - 0{,}54| > \frac{\sigma F^{-1}(1-\alpha)}{n^{1/2}}.$$

A un facteur 2 près, cette règle de décision basée sur le niveau de signification P est donc bien équivalente au test précédent.

6.1.4 Discussion intermédiaire

Bien qu'extrêmement simple, l'exemple ci-dessus illustre plusieurs notions très importantes. Tout d'abord, il est essentiel de comprendre que

En toute rigueur, on ne peut que rejeter une hypothèse.

Reprenons les tests ci-dessus : *in fine*, ce que l'on fait est choisir un évènement – ici, l'évènement $\{|\bar{X}_n - 0{,}54| \geq \kappa\}$ – tel que, **si l'hypothèse était vraie**, alors la probabilité de cet évènement serait faible – ici, fixée à 5%. Ainsi, si l'on observe cet évènement, cela veut dire que **l'hypothèse n'explique pas bien les données** et il est naturel de la rejeter.

A l'inverse, si l'évènement a lieu, on ne peut que conclure que la probabilité de cet évènement sous l'hypothèse n'est pas très faible : néanmoins, de nombreuses autres hypothèses pourraient avoir la même propriété et il ne s'agit donc pas d'une preuve "positive" ! Le raisonnement est similaire à celui qui consisterait à dire que ça n'est pas parce que personne n'a jamais observé de fantômes qu'ils n'existent pas. A l'inverse, une observation de fantômes prouverait automatiquement leur existence.

Cette discussion montre l'importance de l'hypothèse testée : en effet, si les données ne permettent pas de la rejeter, alors on sera amenés à l'accepter et il vaut donc mieux choisir une hypothèse en laquelle on a déjà confiance puisqu'on n'a pas vraiment apporté de preuve de sa validité : on aura seulement montré qu'elle n'était pas incompatible avec les observations. Une manière de pallier ce problème est de contrôler la **puissance du test**, notion que nous introduisons maintenant toujours sur l'exemple de la pièce de monnaie.

6.1.5 Risques de première et deuxième espèce, puissance d'un test

6.1.5.1 Introduction au risque de deuxième espèce

Soit θ^* le vrai biais de la pièce de monnaie : dans l'exemple ci-dessus, on a testé l'hypothèse $\theta^* = 0{,}54$ et, si cette hypothèse était rejetée, on aurait adopté l'hypothèse $\theta^* \neq 0{,}54$. L'hypothèse que l'on teste est appelée **hypothèse nulle** et notée H_0, et l'autre hypothèse est appelée **hypothèse alternative** et est notée H_1. Dans l'exemple ci-dessus on a donc

$H_0 : \theta^* = 0{,}54\,;$

$H_1 : \theta^* \neq 0{,}54.$

On a considéré des tests de la forme

$$\text{Rejeter } H_0 \iff |\bar{X}_n - 0{,}54| \geq \kappa$$

où le seuil κ a été fixé de manière à contrôler le risque de se tromper si l'hypothèse H_0 était vraie :

$$\mathbb{P}_{0{,}54}\left(|\bar{X}_n - 0{,}54| \geq \kappa\right) = \alpha$$

avec $\alpha \in [0, 1]$ le risque que l'on est prêts à prendre. De manière générale, α est appelé **risque de première espèce** :

Le risque de première espèce est la probabilité de rejeter H_0 alors qu'H_0 est vraie.

Comme expliqué ci-dessus, si l'on rejette H_0 alors pas (ou peu) de doutes : H_0 n'expliquait pas bien les observations et il était raisonnable de la rejeter (le risque de première espèce étant là pour quantifier le "peu de doutes"). Par contre, si l'on vient à accepter H_0, on ne peut être sûrs qu'H_0 est la vraie explication des observations et pour accroître la confiance que l'on a dans H_0, on considèrera le **risque de deuxième espèce β :**

Le risque de deuxième espèce β est la probabilité d'accepter H_0 alors qu'H_1 est vraie.

Un risque de deuxième espéce faible est réconfortant : ainsi, bien qu'on ne puisse pas savoir avec certitude si H_0 est la vraie hypothèse qui explique les données – on a juste prouvé que H_0 était compatible avec les données – au moins on sait que H_1 n'est pas compatible avec les données.

A l'inverse, un risque de deuxième espèce élevé est embêtant : il signifie que H_1 aussi est compatible avec les données et cela ne permet donc pas de trancher entre H_0 et H_1.

De manière générale et dans la mesure du possible, on cherchera donc des tests avec des risques de première et deuxième espèce faibles. De manière équivalente, on cherchera à maximiser la puissance η du test, la puissance étant définie comme la probabilité complémentaire du risque de seconde espèce :

$$\eta = 1 - \beta.$$

6.1.5.2 Le cas d'hypothèse composite

On a défini le risque de deuxième espèce comme la probabilité d'accepter H_0 à tort : ainsi, pour calculer β il faut pouvoir faire des calculs sous H_1 ce qui n'est pas possible si H_1 est une **hypothèse composite**, i.e., une hypothèse où la valeur de θ^* n'est pas uniquement déterminée (par opposition à une **hypothèse simple** de la forme $H_0 : \theta^* = 0{,}54$). Si Θ_1 est l'ensemble des valeurs prises par H_1, i.e., $H_1 : \theta^* \in \Theta_1$, alors le risque de deuxième espèce est en fait la fonction

$$\beta : \theta \in \Theta_1 \mapsto \mathbb{P}_\theta \left(\text{Accepter } H_0\right)$$

et la puissance du test est aussi une fonction, définie par $\eta(\theta) = 1 - \beta(\theta)$ pour $\theta \in \Theta_1$. Dans l'exemple de la pièce de monnaie où $\Theta_0 = \{0{,}54\}$ et $\Theta_1 = \Theta \setminus \Theta_0 = [0,1] \setminus \{0{,}54\}$, on peut calculer la fonction puissance, et la courbe obtenue est donnée en Figure 6.1. Pour effectuer le calcul, on note que, par définition

$$\beta(\theta) = \mathbb{P}_\theta \left(0{,}54 - \frac{\sigma}{n^{1/2}} F^{-1}(1-\alpha/2) \leq \bar{X}_n \leq 0{,}54 + \frac{\sigma}{n^{1/2}} F^{-1}(1-\alpha/2)\right).$$

On utilise ensuite le fait que $n^{1/2}(\bar{X}_n - \theta)/\sigma(\theta)$ sous $\mathbb{P}_\theta$ et avec $\sigma(\theta) = (\theta(1-\theta))^{1/2}$ suit approximativement une loi normale pour obtenir

$$\beta(\theta) \approx F\left(\frac{n^{1/2}}{\sigma(\theta)}(0{,}54 - \theta) + \frac{\sigma}{\sigma(\theta)} F^{-1}(1-\alpha/2)\right) - F\left(\frac{n^{1/2}}{\sigma(\theta)}(0{,}54 - \theta) - \frac{\sigma}{\sigma(\theta)} F^{-1}(1-\alpha/2)\right).$$

On voit donc que la puissance est d'autant mauvaise que θ est proche de 0,54, ce qui est tout à fait normal : il est par exemple très difficile de distinguer les hypothèses $\theta^* = 0{,}54$ et $\theta^* = 0{,}55$!

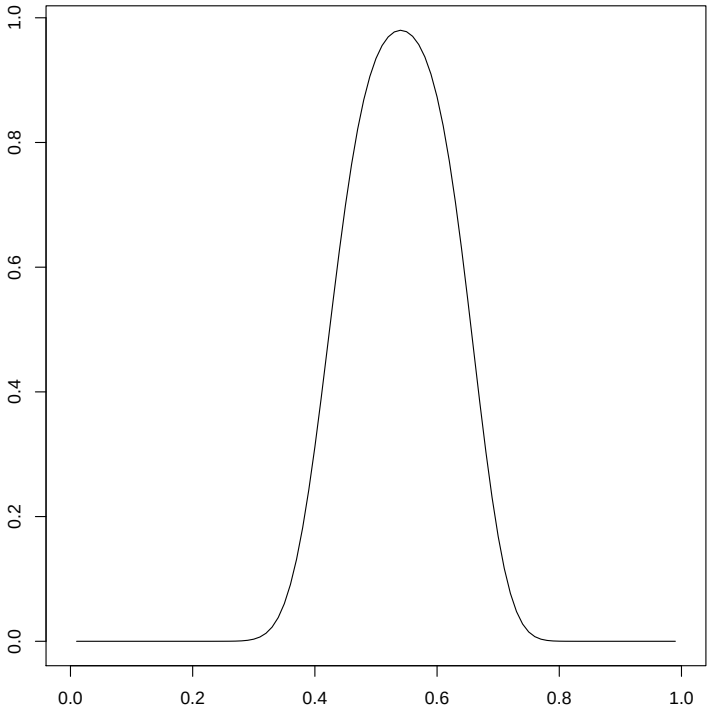


FIGURE 6.1 – Risque de deuxième espèce $\theta \in [0,1] \setminus \{0{,}54\} \mapsto \beta(\theta)$ dans le cas discuté en Section 6.1.5.2.

6.1.6 Hypothèse nulle et hypothèse alternative

L'exemple discuté jusqu'à maintenant consistait en

$H_0 : \theta^* = 0{,}54\,;$

$H_1 : \theta^* \neq 0{,}54\,;$

et le test proposé était de la forme

$$\text{Rejeter } H_0 \iff |\bar{X}_n - 0{,}54| \geq \kappa$$

avec κ choisi de telle sorte à fixer le risque de première espèce $\mathbb{P}_{0{,}54}(\text{Rejeter } H_0) = \alpha$. Comme déjà mentionné précédemment, les hypothèses H_0 et H_1 ne jouent pas un rôle symétrique. On voit que si l'on change H_0, par exemple on considère $H_0 : \theta^* = \theta_0$ alors le test change et devient

$$\text{Rejeter } H_0 \iff |\bar{X}_n - \theta_0| \geq \kappa$$

avec κ tel que $\mathbb{P}_{\theta_0}(|\bar{X}_n - \theta_0| \geq \kappa) = \alpha$. Il semble donc à première vue que l'hypothèse H_1 ne joue pas de rôle ce qui n'est en fait pas le cas : H_1 joue un rôle sur la forme du test. En effet, au lieu de $H_1 : \theta^* \neq \theta_0$ considérons par exemple $H_1 : \theta^* > \theta_1$ pour un certain $\theta_1 > \theta_0$: alors on ne rejetera H_0 pour l'hypothèse alternative H_1 que si l'estimateur $\bar{X}_n$ de θ^* est grand. Ainsi, le test prend la forme

$$\text{Rejeter } H_0 \iff \bar{X}_n \geq \kappa$$

avec κ tel que $\mathbb{P}_{\theta_0}(\bar{X}_n \geq \kappa) = \alpha$. On remarque dans cet exemple que H_1 joue un rôle sur la forme du test, mais que même pour H_1 de la forme $H_1 : \theta^* > \theta_1$ la valeur de θ_1 ne joue pas de rôle sur la détermination du paramètre de seuil κ qui par définition ne dépend que de θ_0 (ou plus généralement Θ_0) et α. En revanche, θ_1 joue un rôle sur la puissance du test :

intuitivement, plus θ_1 est grand et plus il est facile de discriminer entre H_0 et H_1, et plus le test est puissant.

6.2 Résultats généraux

6.2.1 Définitions

Le cadre général des tests d'hypothèse est le suivant. On considère un modèle paramétrique $\{\mathbb{P}_\theta : \theta \in \Theta\}$ et deux sous-ensembles $\Theta_0, \Theta_1 \subset \Theta$ que l'on suppose disjoints : $\Theta_0 \cap \Theta_1 = \emptyset$. Par contre, on ne supppose pas nécessairement que Θ_0 et Θ_1 forment une partition de Θ, i.e., on peut avoir $\Theta_0 \cup \Theta_1 \subset \Theta$ au sens strict. On définit :

H_0 (hypothèse nulle) : $\theta^* \in \Theta_0$;

H_1 (hypothèse alternative) : $\theta^* \in \Theta_1$.

Une hypothèse est dite **simple** si elle est réduite à un singleton, et **composite** ou **multiple** sinon. On observe un échantillon $(X_1, \ldots, X_n)$ de taille n que l'on souhaite utiliser pour décider de rejeter H_0 ou non [1]. On considérera des tests d'hypothèse de la forme suivante :

$$\text{Rejeter } H_0 \iff T_n = T(X_1, \ldots, X_n) \in \Lambda$$

pour une certaine statistique à valeur réelle T appelée **statistique de test** et un certain ensemble $\Lambda \subset \mathbb{R}$, la plupart du temps un intervalle ou une union d'intervalles. La région

$$W = T^{-1}(\Lambda) = \{\mathbf{x} \in \mathbb{R}^n : T(\mathbf{x}) \in \Lambda\}$$

est appelée **région critique** ou **région de rejet** du test : en effet, le test précédent se réécrit sous la forme équivalente

$$\text{Rejeter } H_0 \iff (X_1, \ldots, X_n) \in W.$$

De manière générale, les risques de première et de deuxième espèce ainsi que la puissance d'un test sont des fonctions à valeurs dans $[0, 1]$, définies de la manière suivante.

Définition 6.2.1. Le **risque de première espèce** α, le **risque de deuxième espèce** β et la **puissance du test** η sont les fonctions données par

$$\begin{cases} \alpha : \theta \in \Theta_0 \mapsto \mathbb{P}_\theta(\text{Rejeter } H_0) & \text{(risque de première espèce)}, \\ \beta : \theta \in \Theta_1 \mapsto \mathbb{P}_\theta(\text{Accepter } H_0) & \text{(risque de deuxième espèce)}, \\ \eta : \theta \in \Theta_1 \mapsto 1 - \beta(\theta) & \text{(puissance)}. \end{cases}$$

On remarquera que les fonctions α et β et η diffèrent notamment (et principalement) par leur domaine de définition. Ainsi, α et β représentent des probabilités d'erreur mais sous des hypothèses différentes : α est le risque de se tromper si H_0 est vraie (et donc $\theta \in \Theta_0$) et β le risque de se tromper si H_1 est vraie (et donc $\theta \in \Theta_1$).

1. Comme expliqué précédemment, on ne peut en général pas se servir d'un test pour accepter une hypothèse : ne pas rejeter une hypothèse veut dire qu'il n'existe pas suffisamment de preuves statistiques pour la rejeter, ce qui ne veut pas dire que l'hypothèse est vraie ! Il est plus correct de dire que l'hypothèse est compatible avec les données.

Comme expliqué précédemment, on veut concevoir un test qui, en priorité, contrôle le risque de première espèce : en pratique, on choisira Λ tel que

$$\sup_{\theta\in\Theta_0} \alpha(\theta) = \sup_{\theta\in\Theta_0} \mathbb{P}_\theta\left(T_n \in \Lambda\right) = \alpha$$

pour un niveau de risque α fixé. Ensuite, si l'on a le choix entre plusieurs tests ayant le même risque de première espèce, alors on privilégiera celui de puissance maximale, où la puissance du test est définie par

$$\eta = \sup_{\theta\in\Theta_1} \eta(\theta).$$

6.2.2 Cas d'hypothèses simples

Le cas où H_0 et H_1 sont simples est le seul cas traitable en toute généralité. Le test de Neyman–Pearson est alors, à un niveau de risque de première espèce donné, plus puissant que tout autre test. On rappelle que $\mathcal{L}(\theta;\mathbf{x})$ est la fonction de vraisemblance.

Théorème 6.2.1 (Test de Neyman–Pearson)**.** *Supposons que H_0 et H_1 sont simples :*

H_0 *: $\theta^* = \theta_0$;*

H_1 *: $\theta^* = \theta_1$.*

et considérons le test suivant :

$$\text{Rejeter } H_0 \Longleftrightarrow \frac{\mathcal{L}(\theta_1;\mathbf{X}_n)}{\mathcal{L}(\theta_0;\mathbf{X}_n)} > \kappa_\alpha$$

où le seuil κ_α est déterminé par

$$\mathbb{P}_{\theta_0}\left(\text{Rejeter } H_0\right) = \alpha.$$

Alors la puissance de ce test est meilleure que la puissance de n'importe quel autre test de même risque de première espèce α.

Démonstration. Soit $W = \{\mathbf{x} : \mathcal{L}(\theta_1;\mathbf{x}) > k_\alpha \mathcal{L}(\theta_0;\mathbf{x})\}$ la région critique du test de Neyman–Pearson et W' la région critique d'un autre test de niveau α : le but de la preuve est de montrer que la puissance du test de Neyman–Pearson est plus élevée que la puissance du test de région critique W', i.e., que

$$\mathbb{P}_{\theta_1}(\mathbf{X}_n \in W') \leq \mathbb{P}_{\theta_1}(\mathbf{X}_n \in W) \Longleftrightarrow \mathbb{P}_{\theta_1}(\mathbf{X}_n \in W' \setminus W) \leq \mathbb{P}_{\theta_1}(\mathbf{X}_n \in W \setminus W')$$

où l'équivalence est obtenue en retranchant $\mathbb{P}_{\theta_1}(\mathbf{X}_n \in W' \cap W)$ des deux côtés. Tout d'abord, on note que $\mathbb{P}_{\theta_0}(\mathbf{X}_n \in W) = \mathbb{P}_{\theta_0}(\mathbf{X}_n \in W')$ puisque ces deux probabilités valent α par construction. En retranchant $\mathbb{P}_{\theta_0}(\mathbf{X}_n \in W' \cap W)$, on obtient donc

$$\mathbb{P}_{\theta_0}(\mathbf{X}_n \in W \setminus W') = \mathbb{P}_{\theta_0}(\mathbf{X}_n \in W' \setminus W).$$

Par ailleurs,

$$\mathbb{P}_{\theta_1}(\mathbf{X}_n \in W' \setminus W) = \int \mathbb{1}\left\{\mathbf{x} \in W' \setminus W\right\} \mathcal{L}(\theta_1;\mathbf{x}) \mathrm{d}\mathbf{x}$$

et puisque $\mathcal{L}(\theta_1;\mathbf{x}) \leq k_\alpha \mathcal{L}(\theta_0;\mathbf{x})$ pour $\mathbf{x} \notin W$, on obtient

$$\mathbb{P}_{\theta_1}(\mathbf{X}_n \in W' \setminus W) \leq k_\alpha \int \mathbb{1}\{\mathbf{x} \in W' \setminus W\} \mathcal{L}(\theta_0;\mathbf{x}) \mathrm{d}\mathbf{x} = k_\alpha \mathbb{P}_{\theta_0}(\mathbf{X}_n \in W' \setminus W).$$

On obtient de manière symétrique $\mathbb{P}_{\theta_1}(\mathbf{X}_n \in W \setminus W') > k_\alpha \mathbb{P}_{\theta_0}(\mathbf{X}_n \in W \setminus W')$ et puisque $\mathbb{P}_{\theta_0}(\mathbf{X}_n \in W \setminus W') = \mathbb{P}_{\theta_0}(\mathbf{X}_n \in W' \setminus W)$ cela implique le résultat voulu. ■

6.2.3 Test à base d'intervalles de confiance

Comme on l'a vu en introduction, il est naturel d'utiliser des intervalles de confiance pour construire des tests statistiques. Plus précisément, si Λ est un intervalle de confiance de θ au niveau $1-\alpha$, alors par définition le test

$$\text{Rejeter } H_0 \iff \theta \notin \Lambda$$

est un test de risque de première espèce α : en effet, pour $\theta \in \Theta_0$ on a

$$\mathbb{P}_\theta(\text{Rejeter } H_0) = \mathbb{P}_\theta(\theta \notin \Lambda)$$

qui par définition de l'intervalle de confiance vaut $1-(1-\alpha) = \alpha$. On peut ainsi revisiter les quatre Propositions 5.7.1, 5.7.2, 5.7.3 et 5.7.4 qui donnent autant de tests statistiques pour tester la moyenne à variance connue, la variance à moyenne connue, la variance à moyenne inconnue et la moyenne à variance inconnue, respectivement. Dans les quatre sections qui suivent, on se place donc dans les mêmes conditions que dans la Section 5.7.2, i.e., on considère le modèle gaussien.

6.2.3.1 Test pour la moyenne à variance connue

On fixe $\sigma \in]0,\infty[$ et on considère le modèle paramétrique $\{\mathbb{P}_m : m \in \mathbb{R}\}$ avec $\mathbb{P}_m$ la loi normale de moyenne m et de variance σ^2. La Proposition 5.7.1 montre alors que pour tout a^+, a^- qui satisfont $F(a^+) - F(-a^-) = 1-\alpha$ où F est la fonction de répartition de la loi normale standard, le test

$$\text{Rejeter } H_0 \iff \theta_0 \notin \left[\bar{X}_n - \frac{\sigma a^-}{n^{1/2}}, \bar{X}_n + \frac{\sigma a^+}{n^{1/2}}\right]$$

est un test au niveau α. On a donc le choix sur les paramètres a^+, a^- qui sera guidé par la forme de l'hypothèse alternative H_1. On considère trois cas :

Premier cas : $H_1 : \theta^* \neq \theta_0$. On considère alors l'intervalle bilatéral symétrique donné par $a^+ = a^- = F^{-1}(1-\alpha/2)$.

Deuxième cas : $H_1 : \theta^* > \theta_1$ avec $\theta_1 > \theta_0$. On considère alors un test de la forme

$$\text{Rejeter } H_0 \iff \bar{X}_n > \kappa$$

ce qui revient à considérer l'intervalle unilatéral à droite, i.e., $a^+ = \infty$ et $a^- = F^{-1}(1-\alpha)$;

Troisième cas : $H_1 : \theta^* < \theta_1$ avec $\theta_1 < \theta_0$. On fait un raisonnement symétrique au cas précédent et on considère l'intervalle unilatéral à gauche.

6.2.3.2 Intervalle de confiance pour la variance à moyenne connue

On fixe maintenant $m \in \mathbb{R}$ et on considère le modèle paramétrique $\{\mathbb{P}_\sigma : \sigma \in \mathbb{R}_+\}$ avec $\mathbb{P}_\sigma$ la loi normale de moyenne m fixée et de variance σ^2. La Proposition 5.7.2 montre alors que pour tout a^+, a^- satisfaisant $G_n(a^+) - G_n(a^-) = 1-\alpha$ avec G_n la fonction de répartition de la loi du χ^2 à n degrés de liberté, le test

$$\text{Rejeter } H_0 \iff \theta_0 \notin \left[\frac{n\hat{\sigma}_n^2}{a^+}, \frac{n\hat{\sigma}_n^2}{a^-}\right]$$

est un test au niveau α.

Premier cas : $H_1 : \theta^* \neq \theta_0$. On considère alors l'intervalle bilatéral symétrique donné par le choix $a^- = G_n^{-1}(\alpha/2)$ et $a^+ = G_n^{-1}(1-\alpha/2)$.

Deuxième cas : $H_1 : \theta^* > \theta_1$ avec $\theta_1 > \theta_0$. On considère alors l'intervalle unilatéral à droite.

Troisième cas : $H_1 : \theta^* < \theta_1$ avec $\theta_1 < \theta_0$. On considère alors l'intervalle unilatéral à gauche.

6.2.3.3 Intervalle de confiance pour la variance à moyenne inconnue

On considère le même modèle paramétrique que précédemment mais on suppose maintenant que la moyenne est inconnue. La Proposition 5.7.3 montre alors que pour tout $a^+ > a^- > 0$ satisfaisant $G_{n-1}(a^+) - G_{n-1}(a^-) = 1-\alpha$, le test

$$\text{Rejeter } H_0 \iff \theta_0 \notin \left[\frac{(n-1)S_{n-1}^2}{a^+}, \frac{(n-1)S_{n-1}^2}{a^-}\right]$$

est un test au niveau α.

Premier cas : $H_1 : \theta^* \neq \theta_0$. On considère alors l'intervalle bilatéral symétrique donné par le choix $a^- = G_{n-1}^{-1}(\alpha/2)$ et $a^+ = G_{n-1}^{-1}(1-\alpha/2)$.

Deuxième cas : $H_1 : \theta^* > \theta_1$ avec $\theta_1 > \theta_0$. On considère alors l'intervalle unilatéral à droite.

Troisième cas : $H_1 : \theta^* < \theta_1$ avec $\theta_1 < \theta_0$. On considère alors l'intervalle unilatéral à gauche.

6.2.3.4 Test de Student : intervalle de confiance pour la moyenne à variance inconnue

On conclut par la construction d'intervalles de confiance pour la moyenne lorsque la variance est inconnue : on considère donc le modèle paramétrique $\{\mathbb{P}_{m,\sigma} : m \in \mathbb{R}, \sigma \in \mathbb{R}_+\}$ avec $\mathbb{P}_{m,\sigma}$ la loi normale de moyenne m et de variance σ^2 (on sort donc du cadre de la dimension un considéré dans le reste du chapitre). La Proposition 5.7.4 montre alors que pour tout $a^+, a^- > 0$ satisfaisant $H_{n-1}(a^+) - H_{n-1}(-a^-) = 1-\alpha$ avec H_n la fonction de répartition de la loi de Student à n degrés de liberté, le test

$$\text{Rejeter } H_0 \iff \theta_0 \notin \left[\bar{X}_n - \frac{S_{n-1}a^-}{n^{1/2}}, \bar{X}_n + \frac{S_{n-1}a^+}{n^{1/2}}\right]$$

est un test au niveau α.

Premier cas : $H_1 : \theta^* \neq \theta_0$. On considère alors l'intervalle bilatéral symétrique donné par le choix $a^- = H_{n-1}^{-1}(\alpha/2)$ et $a^+ = H_{n-1}^{-1}(1-\alpha/2)$.

Deuxième cas : $H_1 : \theta^* > \theta_1$ avec $\theta_1 > \theta_0$. On considère alors l'intervalle unilatéral à droite.

Troisième cas : $H_1 : \theta^* < \theta_1$ avec $\theta_1 < \theta_0$. On considère alors l'intervalle unilatéral à gauche.

Il s'agit d'un test extrêment utile en pratique, qui s'appelle le **test de Student**.

6.3 Fiche de synthèse

Les *tests statistiques* sont utilisés lorsqu'on cherche à savoir si une certaine hypothèse relative à la population est compatible avec l'information disponible à partir de l'échantillon.

On considère un modèle paramétrique $\{\mathbb{P}_\theta \ ; \ \theta \in \Theta\}$ et deux sous-ensembles disjoints Θ_0 et Θ_1 de Θ.
$\rightarrow$ Une hypothèse H concernant le paramètre θ^* est énoncée et traduite sous la forme de 2 propositions contradictoires.

- H_0 dite *hypothèse nulle* : $\theta^* \in \Theta_0$ (test simple si $\Theta = \{\theta_0\}$, composite sinon)
- H_1 dite *hypothèse alternative* : $\theta^* \in \Theta_1$. Si H_0 est $\theta^* = \theta_0$, H_1 peut être $\theta^* \neq \theta_0$ (test bilatéral) ou bien $\theta^* > \theta_0$ ou $\theta^* < \theta_0$ (test unilatéral).

$\rightarrow$ On utilise un échantillon $\mathbf{X}_n = (X_1, \cdots, X_n)$ et une statistique de test $T_n = f(\mathbf{X}_n)$ (par exemple $\mathbf{X}_n$) dont la loi sous H_0 est parfaitement connue (T_n n'est pas unique !).

Règle de décision, risques d'erreur : on fixe $\alpha \in]0,1[$ (souvent $0,05$ ou $0,01$)
$\rightarrow$ On a alors $\alpha = \mathbb{P}_{H_0}(\text{rejeter } H_0)$ *risque de 1er type* : risque de rejeter à tort H_0, contrôlé.
Pour un test composite, $\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\text{rejeter } H_0)$.
Si $\Theta_0 = \{\theta_0\}$, on détermine Λ (souvent union d'intervalles) tel que $\alpha = \mathbb{P}_{\theta_0}(\text{rejeter } H_0) = \mathbb{P}_{\theta_0}(T_n \in \Lambda)$. L'ensemble $W = \{\mathbf{x} \in \mathbb{R}^n \ ; \ T_n(\mathbf{x}) \in \Lambda\}$ est la *région critique*.
$\rightarrow$ $\beta = \mathbb{P}_{H_1}(\text{accepter } H_0)$ *risque de 2ème type* : risque d'accepter à tort H_0, non contrôlé.
$\rightarrow$ Pour $\theta \in \Theta_1$, $\eta(\theta) = 1 - \beta(\theta)$ *puissance du test* : si on a $H_0 : \theta^* = \theta_0$ et $H_1 : \theta^* = \theta_1$, le test de plus grande puissance parmi les tests de risque de 1er type α est le test de Neyman-Pearson qui consiste à rejeter H_0 ssi $\frac{\mathcal{L}(\theta_1, \mathbf{X}_n)}{\mathcal{L}(\theta_0, \mathbf{X}_n)} > \kappa_\alpha$ où κ_α est déterminé par $\mathbb{P}_{\theta_0}(\text{rejeter } H_0) = \alpha$.

Pratique des tests : le tableau suivant envisage des cas fréquents.

paramètre à tester		statistique de test	loi
moyenne m	variance σ^2 connue	$\frac{\sqrt{n}(\overline{X}_n - m)}{\sigma}$	$\mathcal{N}(0;1)$
	variance inconnue	$\frac{\sqrt{n}(\overline{X}_n - m)}{S_{n-1}}$	Student à $n-1$ *d.d.l.*
variance σ^2	moyenne m connue	$\frac{1}{\sigma^2}\sum_{k=1}^{n}(X_k - m)^2$	χ^2 à n *d.d.l.*
	moyenne inconnue	$\frac{(n-1)S_{n-1}^2}{\sigma^2}$	χ^2 à $n-1$ *d.d.l.*

Lois de probabilités utilisées dans les tests

• Loi du χ^2 à n degrés de liberté (d.d.l.) : de densité $x \mapsto \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2} \mathbb{1}\{x > 0\}$. C'est aussi la loi Gamma de paramètres $n/2$ et $1/2$.
Propriété : Si X_k sont des variables aléatoires indépendantes de loi normale $\mathcal{N}(m;\sigma^2)$, alors $\sum_{k=1}^{n}\left(\frac{X_k - m}{\sigma}\right)^2$ suit la loi du χ^2 à n d.d.l.

• Loi de Student à n d.d.l. : de densité $x \mapsto \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$.
Propriété : C'est la loi de $\frac{X}{\sqrt{Z_n/n}}$ où X suit la loi normale $\mathcal{N}(0;1)$ et est indépendante de Z_n qui suit la loi du χ^2 à n d.d.l.

• Loi de Fisher-Snedecor de paramètre (p,q) : de densité $x \mapsto \frac{\left(\frac{px}{px+q}\right)^{p/2}\left(1 - \frac{px}{px+q}\right)^{q/2}}{xB(p/2, q/2)} \mathbb{1}\{x > 0\}$
Propriété : C'est la loi de $\frac{Z_p/p}{Z_q/q}$ où Z_p et Z_q sont indépendantes et suivent des lois du χ^2 à p et q d.d.l. respectivement.

Théorème 4.6.2 : Si $X_1, \cdots, X_n$ sont des variables aléatoires indépendantes de loi normale $\mathcal{N}(m;\sigma^2)$, alors les variables aléatoires

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k \text{ et } S_{n-1}^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2$$

sont indépendantes.

Par ailleurs :

- $\overline{X}_n$ suit la loi normale $\mathcal{N}(m, \sigma^2/n)$;
- $(n-1)S_{n-1}^2/\sigma^2$ suit la loi du χ^2 à $n-1$ d.d.l. ;
- $\frac{\sqrt{n}(\overline{X}_n - m)}{S_{n-1}}$ suit la loi de Student à $n-1$ d.d.l.

6.4 Exercices

Les exercices précédés d'une flèche ↪ sont des exercices d'application directs du cours.

↪ **Exercice 6.1** *(Risques de première et de deuxième espèce)*

1. On considère le modèle gaussien à variance $\sigma^2 = 4$ connue et l'on souhaite tester l'hypothèse nulle $H_0 : \theta = 1$ contre l'hypothèse alternative $H_1 : \theta = 2$. Quelle est la taille d'échantillons minimale qui garantisse un risque de première espèce $\alpha = 3\%$ et une puissance $\eta = 80\%$?

2. On considère le même modèle qu'à la question précédente. Parmi les hypothèses alternatives suivantes, pour lesquelles peut-on tester l'hypothèse nulle $H_0 : \theta = 2$ avec un risque de $\alpha = 3\%$ et une puissance $\eta = 85\%$? Commentez les résultats obtenus.

Cas a. $H_1 : \theta = 3$

Cas b. $H_1 : \theta = 1{,}999$

Cas c. $H_1 : \theta < 2$

3. On considère un test statistique de la forme

$$\text{Rejeter } H_0 \iff \left|\hat{\theta}_n - \theta_0\right| > \kappa$$

pour tester deux hypothèses simples $H_0 : \theta^* = \theta_0$ contre $H_1 : \theta^* = \theta_1$. Dans quel sens les risques de première et de deuxième espèce varient-ils lorsque κ varie ? Quel est le problème de fixer un risque de première espèce très faible ?

↪ **Exercice 6.2**

On considère un échantillon $X_1, \dots, X_n$ de variables aléatoires i.i.d. tirées sur la loi uniforme sur $[0, \theta]$. On souhaite tester l'hypothèse nulle $H_0 : \theta = 1$ contre l'hypothèse alternative $H_1 : \theta = 1 + \varepsilon$. On considère les statistiques $M_n = \max_{k=1,\dots,n} X_k$ et $\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$ que l'on souhaite utiliser comme statistique de test.

1. Expliquez pourquoi les deux tests suivants sont raisonnables :

$$\text{Accepter } H_0 \iff \bar{X}_n < \kappa^X \quad \text{et} \quad \text{Accepter } H_0 \iff M_n < \kappa^M.$$

A-t-on $\kappa^X < 1/2$ ou $\kappa^X > 1/2$? et $\kappa^M < 1$ ou $\kappa^M > 1$?

2. Quelle est la limite en loi de $\sqrt{n}(\bar{X}_n - \frac{1}{2})$?

3. Utilisez la question précédente pour montrer que si l'on se fixe un risque de première espèce α donné, alors κ^X et κ^M doivent satisfaire

$$\kappa^X \approx \frac{1}{2} + \frac{F^{-1}(1-\alpha)}{2\sqrt{3n}} \quad \text{et} \quad \kappa^M = (1-\alpha)^{1/n}.$$

4. Montrez que, sous l'approximation normale précédente et en utilisant donc les valeurs de κ^X et κ^M de la question précédente, les puissances η^X et η^M associées à ces tests sont données par

$$\eta^X \approx 1 - F\left(-\frac{\varepsilon\sqrt{3n}}{1+\varepsilon} + (1+\varepsilon)^{-1}F^{-1}(1-\alpha)\right) \quad \text{et} \quad \eta^M = 1 - \frac{1-\alpha}{(1+\varepsilon)^n}$$

avec z et z' certaines constantes indépendantes de n.

5. En utilisant l'approximation (fausse mais instructive) $F(x) \approx 1 - \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, justifiez les approximations

$$-\log(1-\eta^X) \approx \frac{3\varepsilon^2 n}{2} \quad \text{et} \quad -\log(1-\eta^M) \approx n\varepsilon.$$

Quel test utiliseriez-vous lorsque ε est petit ?

6. Peut-on utiliser le test de Neyman–Pearson ? Pourquoi ?

↪ **Exercice 6.3**

On distingue deux types de plantes à graines : les plantes qui disséminent leurs graines très localement avec une grande probabilité et loin avec une faible probabilité (type 0) et les plantes qui disséminent leurs graines localement (type 1). La loi de la position relative (X, Y) d'une graine par rapport à la plante est :

- pour le type 0, la loi gaussienne centrée réduite de densité

$$p_0(x,y) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}(x^2+y^2)\right);$$

- pour le type 1, la loi uniforme sur $[-d,d]^2$ avec $d=2$, de densité

$$p_1(x,y) = \frac{1}{d^2}\mathbb{1}\left\{x,y\in[-d,d]^2\right\} = \frac{1}{4}\mathbb{1}\left\{x,y\in[-2,2]\right\};$$

La graine d'une plante est observée à la position relative (x,y) par rapport à la plante. On souhaite savoir s'il s'agit d'une plante de type 0. Dans les deux premières questions on supposera que H_0 est l'hypothèse "la plante est de type 0" et H_1 l'hypothèse "la plante est de type 1".

1. Décrivez le test de Neyman–Pearson au niveau de risque de première espèce α ainsi que sa région critique.

2. Soit G_2 la fonction de répartition de la loi du χ^2 à 2 degrés de liberté : montrez que le risque de première espèce du test

$$\text{Rejeter } H_0 \iff |X|,|Y| \le d, X^2+Y^2 > G_2^{-1}(1-\alpha)$$

est majoré par α.

3. On donne $G_2^{-1}(0{,}95) \approx 6$ et $G_2^{-1}(0{,}99) \approx 9$. Dessinez la région critique du test de la question 2 pour $\alpha = 5\%$ et $\alpha = 1\%$ et commentez les résultats obtenus.

↪ **Exercice 6.4** *(Comparaison de moyenne et de variance)*

On considère deux échantillons $X_1,\dots,X_{n_X}$ et $Y_1,\dots,Y_{n_Y}$: toutes les variables aléatoires sont indépendantes, et on suppose en outre que les X_i suivent une loi normale de paramètre (m_X,σ_X) et que les Y_i suivent une loi normale de paramètre (m_Y,σ_Y).
On définit

$$\bar{X} = \frac{1}{n_X}\sum_{k=1}^{n_X} X_k \quad\text{et}\quad S_X^2 = \frac{1}{n_X-1}\sum_{k=1}^{n_X}\left(X_k-\bar{X}\right)^2$$

ainsi que

$$\bar{Y} = \frac{1}{n_Y}\sum_{k=1}^{n_Y} Y_k \quad\text{et}\quad S_Y^2 = \frac{1}{n_Y-1}\sum_{k=1}^{n_Y}\left(Y_k-\bar{Y}\right)^2.$$

1. Quelle est la loi de $(S_X^2/\sigma_X^2)/(S_Y^2/\sigma_Y^2)$? Quelle est la loi de S_X^2/S_Y^2 sous H_0 ? (Reportez-vous au paragraphe "Lois de probabilités utilisées dans les tests" de la fiche de synthèse de ce chapitre)

2. Déduisez de la question précédente un test pour les deux cas suivants : $H_0 : \sigma_X = \sigma_Y$ et $H_1 : \sigma_X > \sigma_Y$.

3. Même question mais pour tester $H_0 : \sigma_X = \sigma_Y$ contre $H_1 : \sigma_X \neq \sigma_Y$, puis commentez la différence entre les deux tests.

On suppose que le test précédent a mené à la conclusion d'égalité des variances $\sigma_X = \sigma_Y = \sigma$ et l'on considérera donc dorénavant cette hypothèse comme étant satisfaite. On définit $n = n_X + n_Y$ et

$$S^2 = \frac{1}{n-2}\left(\sum_{k=1}^{n_X}\left(X_k-\bar{X}\right)^2 + \sum_{k=1}^{n_Y}\left(Y_k-\bar{Y}\right)^2\right).$$

4. Quelle est la loi de $(n-2)S^2/\sigma^2$ et de $\bar{X}-\bar{Y}$?

5. Montrez que S est indépendante de $\bar{X}$ et $\bar{Y}$ et déduisez-en la loi de

$$\frac{\bar{X}-\bar{Y}-(m_X-m_Y)}{S\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}}.$$

6. Déduisez un test de Student pour tester l'égalité des moyennes, i.e., pour tester l'hypothèse $H_0 : m_X = m_Y$ contre $H_1 : m_X \neq m_Y$.

↪ Problème 6.5

Dans les conditions usuelles d'emploi, la durée de vie des pièces produites par une machine est modélisée par une loi gaussienne $\mathcal{N}(\mu_0, \sigma^2)$, avec $\mu_0 = 120$ heures et $\sigma = 19{,}4$ heures. Le représentant d'un fournisseur propose un nouveau modèle de machine, en promettant un gain sur la durée de vie des pièces produites de 5% en moyenne, pour un écart-type identique σ.
On décide de tester le nouveau modèle de machine sur un échantillon de $n = 64$ pièces produites. On note $(X_i, i \in \{1, ..., n\})$ les durées de vie des n pièces produites par le nouveau modèle de machine.

1. Quelle est la loi de $(X_i, i \in \{1, ..., n\})$?

2. Soit μ la durée de vie moyenne des pièces produites par le nouveau modèle de machine. Donner un estimateur sans biais de μ. Identifier la loi de cet estimateur.

3. On ne souhaite pas changer de modèle si le nouveau n'est pas plus performant que l'ancien. Plus précisément, on souhaite que la probabilité d'adopter à tort le nouveau modèle ne dépasse pas le seuil de $\alpha = 0{,}05$. Quelle est alors la procédure de décision construite à partir de l'estimateur de μ ? Les 64 pièces testées ont eu une durée de vie moyenne égale à 123,5 heures. Conclusion.

4. Evaluez le risque que cette procédure vous fasse rejeter le nouveau modèle si l'annonce du représentant est exacte.

Le représentant conteste cette procédure, prétextant qu'il vaut mieux partir de l'hypothèse H_0' , selon laquelle le gain de performance moyen est réellement de 5%, tout en conservant le même seuil α pour ce test.

5. Quelle est alors la procédure de décision ? Quel est le risque de l'acheteur ? Quel est le résultat de cette procédure au vu des observations faites. Conclusion.

6. Quelle procédure peut-on proposer pour égaliser les risques de l'acheteur et du vendeur ? Quel est alors ce risque ?

Annexe A

Correction des exercices

A.1 Exercices du Chapitre 1

↪ **Exercice 1.1** *(Manipulation de l'espérance conditionnelle)*

1. Supposons que $Y_1 \geq Y_2$: alors $Y_1 \mathbb{1}\{X = x\} \geq Y_2 \mathbb{1}\{X = x\}$ et la Proposition 1.3.7 donne $\mathbb{E}(Y_1; X = x) \geq \mathbb{E}(Y_2; X = x)$. Puisque $\mathbb{E}(Y_i \mid X = x) = \mathbb{E}(Y_i; X = x)/\mathbb{P}(X = x)$ (Théorème 1.6.2), il vient $\mathbb{E}(Y_1 \mid X = x) \geq \mathbb{E}(Y_2 \mid X = x)$ après division par $\mathbb{P}(X = x)$. Soit h_i la fonction telle que $\mathbb{E}(Y_i \mid X) = h_i(X)$: on vient de prouver $h_1(x) \geq h_2(x)$ pour tout x, et donc on a bien $h_1(X) \geq h_2(X)$.
Les relations $\mathbb{E}(aY_1 + bY_2 \mid X) = a\mathbb{E}(Y_1 \mid X) + b\mathbb{E}(Y_2 \mid X)$ et $|\mathbb{E}(Y \mid X)| \leq \mathbb{E}(|Y| \mid X)$ se prouvent de la même manière.
Pour $\mathbb{E}(\varphi(X)Y \mid X) = \varphi(X)\mathbb{E}(Y \mid X)$, il faut aussi utiliser le fait que $\mathbb{E}(\varphi(X)Y \mid X = x) = \varphi(x)\mathbb{E}(Y \mid X = x)$: pour cela, on utilisera le Théorème 1.6.2 et le fait que $\varphi(X)\mathbb{1}\{X = x\} = \varphi(x)\mathbb{1}\{X = x\}$.

2. On prouve le résultat pour $\varphi(X) \geq 0$. Soit $x \in X(\Omega)$ avec $\mathbb{P}(X = x) > 0$, $h(x) = \mathbb{E}(\varphi(Y) \mid X = x)$ et $h_y(x) = \mathbb{P}(Y = y \mid X = x)$: alors

$$\begin{aligned}
h(x) &= \frac{1}{\mathbb{P}(X = x)} \mathbb{E}(\varphi(Y); X = x) && \text{(Théorème 1.6.2)} \\
&= \frac{1}{\mathbb{P}(X = x)} \sum_{x', y} \varphi(y) \mathbb{1}\{x' = x\} \mathbb{P}(Y = y, X = x') && \text{(Théorème 1.3.6 appliqué à } (X, Y)) \\
&= \frac{1}{\mathbb{P}(X = x)} \sum_{y} \varphi(y) \mathbb{P}(Y = y, X = x) \\
&= \sum_{y} \varphi(y) h_y(x) && \text{(Définition 1.6.1).}
\end{aligned}$$

Ainsi, les fonctions h et $\sum_y \varphi(y)h_y$ sont identiques, et donc $h(X) = \sum_y \varphi(y)h_y(X)$ ce qui est exactement ce qu'il fallait montrer.

3. On a $\mathbb{P}(X' \leq X \mid X = x) = \mathbb{P}(X' \leq x \mid X = x)$ et puisque X et X' sont indépendantes, $\mathbb{P}(X' \leq x \mid X = x) = \mathbb{P}(X' \leq x)$ qui vaut $F_X(x)$ puisque X' a la même loi que X. On vient donc de prouver que $\mathbb{P}(X' \leq X \mid X) = F_X(X)$.

4. On montre la formule générale : on a

$$\mathbb{E}(Y \mid X = x)\mathbb{1}\{X = x\} = h(x)\mathbb{1}\{X = x\} = h(X)\mathbb{1}\{X = x\}$$

et donc

$$\sum_{x \in X(\Omega)} \mathbb{E}(Y \mid X = x)\mathbb{1}\{X = x\} = \sum_{x \in X(\Omega)} h(X)\mathbb{1}\{X = x\} = h(X) \sum_{x \in X(\Omega)} \mathbb{1}\{X = x\} = h(X).$$

Pour $X = \xi_A$ cela donne $\mathbb{E}(Y \mid \xi_A) = \mathbb{1}\{\xi_A = 1\}\mathbb{E}(Y \mid \xi_A = 1) + \mathbb{1}\{\xi_A = 0\}\mathbb{E}(Y \mid \xi_A = 0)$ ce qui donne bien le résultat puisque $\{\xi_A = 1\} = A$, $\{\xi_A = 0\} = A^c$ et $\mathbb{1}\{\xi_A = 1\} = \xi_A$.

5. La Proposition 1.3.8 donne

$$\mathbb{E}\left(\sum_{k=1}^{N} X_i \mid N\right) = \mathbb{E}\left(\sum_{k\geq 1} \mathbb{1}\{k \leq N\} X_i \mid N\right) = \sum_{k\geq 1} \mathbb{E}(\mathbb{1}\{k \leq N\} X_i \mid N).$$

Puisque $\mathbb{1}\{k \leq N\}$ est une fonction de N, la Proposition 1.7.2 donne

$$\mathbb{E}(\mathbb{1}\{k \leq N\} X_i \mid N) = \mathbb{1}\{k \leq N\} \mathbb{E}(X_i \mid N)$$

et donc

$$\mathbb{E}\left(\sum_{k=1}^{N} X_i \mid N\right) = \sum_{k\geq 1} \mathbb{1}\{k \leq N\} \mathbb{E}(X_i \mid N) = \sum_{k=1}^{N} \mathbb{E}(X_i \mid N).$$

6. Par des raisonnements similaires à ci-dessus, on voit que

$$\mathbb{E}\left(\prod_{k=1}^{N} X_i \mid N = n\right) = \mathbb{E}\left(\prod_{k=1}^{n} X_i \mid N = n\right)$$

et puisque les X_i sont indépendants et N, la Proposition 1.4.4 donne

$$\mathbb{E}\left(\prod_{k=1}^{n} X_i \mid N = n\right) = \mathbb{E}\left(\prod_{k=1}^{n} X_i\right) = \mathbb{E}(X_1)^n$$

ce qui montre qui prouve le résultat.

↪ **Exercice 1.2** *(Calcul d'espérances conditionnelles)*

1. Pour $k, s \in \mathbb{N}^*$ avec $k \leq s-1$ on a

$$\mathbb{P}(X = k \mid X+Y = s) = \frac{\mathbb{P}(X = k, Y = s-k)}{\mathbb{P}(X+Y = s)} = \frac{(1-p)^{k-1}p(1-p)^{s-k-1}p}{\sum_{j=1}^{s-1}(1-p)^{j-1}p(1-p)^{s-j-1}p} = \frac{1}{s-1},$$

et donc X sachant $X+Y$ est uniformément réparti sur $\{1, \ldots, X+Y-1\}$. Puisque la moyenne de la loi uniforme sur $\{1, \ldots, n\}$ vaut $(n+1)/2$ (cf. Tableau B.1) on en déduit que

$$\mathbb{E}(X \mid X+Y) = \frac{X+Y}{2}.$$

2. Puisque les X_i sont i.i.d., on a $\mathbb{E}(X_i \mid X_1 + \cdots + X_n) = \mathbb{E}(X_1 \mid X_1 + \cdots + X_n)$. Par ailleurs, on a par linéarité de l'espérance conditionnelle

$$\sum_{i=1}^{n} \mathbb{E}(X_i \mid X_1 + \cdots + X_n) = \mathbb{E}\left(\sum_{i=1}^{n} X_i \mid X_1 + \cdots + X_n\right) = \sum_{i=1}^{n} X_i$$

en utilisant la Proposition 1.7.2 pour la dernière égalité, ce qui donne le résultat.

3. Par calcul direct :

$$\mathbb{P}(X+Y = s) = \sum_{k=0}^{s} \mathbb{P}(X = k, Y = s-k) = \sum_{k=0}^{s} e^{-a}\frac{a^k}{k!}e^{-b}\frac{b^{s-k}}{(s-k)!} = e^{-(a+b)}\frac{(a+b)^s}{s!}, \ s \in \mathbb{N},$$

et par la fonction génératrice :

$$\phi_{X+Y}(z) = \phi_X(z)\phi_Y(z) = \exp(-a(1-z))\exp(-b(1-z)) = \exp(-(a+b)(1-z)), \ z \in [-1,1]$$

en utilisant la Proposition 1.5.2 pour la première égalité (la fonction génératrice de la somme de variables indépendantes est égale au produit des fonctions génératices) puis l'expression de la fonction génératrice de

la loi de Poisson pour la seconde égalité (cf. Tableau B.1). On reconnaît bien la fonction génératrice d'une loi de Poisson de paramètre $a+b$, ce qui donne le résultat. Pour $x, s \in \mathbb{N}$ avec $x \leq s$ cela donne

$$\mathbb{P}\left(X = x \mid X+Y = s\right) = \frac{\mathbb{P}(X=x)\mathbb{P}(Y=s-x)}{\mathbb{P}(X+Y=s)} = \frac{e^{-a}\frac{a^x}{x!}e^{-b}\frac{b^{s-x}}{(s-x)!}}{e^{-(a+b)}\frac{(a+b)^s}{s!}} = \binom{s}{x}\left(\frac{a}{a+b}\right)^x\left(\frac{b}{a+b}\right)^{s-x}.$$

La loi conditionnelle de X sachant $X+Y$ est donc une loi binomiale de paramètre $(X+Y, a/(a+b))$ d'où on déduit

$$\mathbb{E}(X \mid X+Y) = \frac{a}{a+b}(X+Y)$$

puisque la moyenne de la loi binomiale de paramètre (n,p) est np, cf. Tableau B.1.

4. Pour $\alpha \in \{1, \ldots, n\}$ on a

$$\mathbb{P}(\max(X,Y) = \min(X,Y) = \alpha) = \mathbb{P}(X = Y = \alpha) = \frac{1}{n^2}$$

et pour $\beta > \alpha$ dans $\{1, \ldots, n\}$,

$$\mathbb{P}(\max(X,Y) = \beta, \min(X,Y) = \alpha) = \mathbb{P}(X=\alpha, Y=\beta) + \mathbb{P}(X=\beta, Y=\alpha) = \frac{2}{n^2}$$

ce qui donne

$$\mathbb{P}(\min(X,Y) = \alpha) = \sum_{\beta \geq \alpha} \mathbb{P}(\max(X,Y) = \beta, \min(X,Y) = \alpha) = \frac{2n - 2\alpha + 1}{n^2}$$

et donc

$$\mathbb{P}(\max(X,Y) = \beta \mid \min(X,Y) = \alpha) = \begin{cases} 0 & \text{si } \beta < \alpha, \\ \dfrac{1}{2n-2\alpha+1} & \text{si } \beta = \alpha, \\ \dfrac{2}{2n-2\alpha+1} & \text{si } \beta > \alpha. \end{cases}$$

et donc

$$\mathbb{P}(\max(X,Y) = \beta \mid \min(X,Y)) = \frac{\mathbb{1}\{\min(X,Y) = \beta\} + 2\mathbb{1}\{\min(X,Y) < \beta\}}{2n - 2\min(X,Y) + 1}.$$

↪ **Exercice 1.3** *(Théorème de la variance totale)*

1. Soit $Z = \mathbb{E}(Y \mid X)$. Pour la première formule, on commence par développer le carré

$$\mathbb{V}\mathrm{ar}(Y \mid X) = \mathbb{E}\left(Y^2 - 2YZ + Z^2 \mid X\right)$$

et puisque l'espérance conditionnelle est linéaire (Proposition 1.7.2) cela donne

$$\mathbb{V}\mathrm{ar}(Y \mid X) = \mathbb{E}\left(Y^2 \mid X\right) - 2\mathbb{E}\left(YZ \mid X\right) + \mathbb{E}\left(Z^2 \mid X\right).$$

Puisque $Z = \mathbb{E}(Y \mid X)$ est par définition une fonction de X, on obtient (Proposition 1.7.2)

$$\mathbb{E}\left(YZ \mid X\right) = Z\mathbb{E}\left(Y \mid X\right) = Z^2$$

ainsi que $\mathbb{E}(Z^2 \mid X) = Z^2$ ce qui donne bien la première formule.
Pour la deuxième formule, on écrit

$$\mathbb{V}\mathrm{ar}(Y) = \mathbb{E}\left(\left(Y - \mathbb{E}(Y)\right)^2\right) = \mathbb{E}\left(\left(Y - Z + Z - \mathbb{E}(Y)\right)^2\right) = A + B + 2C$$

avec

$$A = \mathbb{E}\left(\left(Y - Z\right)^2\right), \; B = \mathbb{E}\left(\left(Z - \mathbb{E}(Y)\right)^2\right)$$

et

$$C = \mathbb{E}\left((Y - Z)\left(Z - \mathbb{E}(Y)\right)\right).$$

Le théorème de l'espérance totale (Théorème 1.7.3) donne

$$A = \mathbb{E}\left[\mathbb{E}((Y - Z)^2 \mid X)\right] = \mathbb{E}(\mathbb{V}\mathrm{ar}(Y \mid X))$$

ainsi que

$$B = \mathbb{E}((Z - \mathbb{E}(Y))^2) = \mathbb{E}((Z - \mathbb{E}(Z))^2) = \mathbb{V}\mathrm{ar}(Z) = \mathbb{V}\mathrm{ar}(\mathbb{E}(Y \mid X))$$

et enfin

$$C = \mathbb{E}\left[\mathbb{E}\left((Y - Z)(Z - \mathbb{E}(Y)) \mid X\right)\right].$$

Puisque $Z - \mathbb{E}(Y)$ est une fonction de X, on a comme précédemment

$$\mathbb{E}\left((Y - Z)(Z - \mathbb{E}(Y)) \mid X\right) = (Z - \mathbb{E}(Y))\mathbb{E}\left(Y - Z \mid X\right)$$

et puisque

$$\mathbb{E}\left(Y - Z \mid X\right) = \mathbb{E}(Y \mid X) - \mathbb{E}(Z \mid X) = Z - Z = 0$$

on obtient $C = 0$, ce qui prouve le résultat.

2. Soit N le nombre de clients dans une semaine et D_i la dépense du $i$ème client, si bien que le chiffre d'affaires R est donné par

$$R = \sum_{k=1}^{N} D_i.$$

Puisque les D_i sont i.i.d. et indépendants de N, on a $\mathbb{E}(D_i \mid N) = \mathbb{E}(D)$ et donc

$$\mathbb{E}(R) = \mathbb{E}\left[\mathbb{E}\left(\sum_{k=1}^{N} D_i \mid N\right)\right] = \mathbb{E}\left[\sum_{k=1}^{N} \mathbb{E}\left(D_i \mid N\right)\right] = \mathbb{E}(D)\mathbb{E}(N) = 4 \times 10^4.$$

En outre, on vient de calculer $\mathbb{E}(R \mid N) = N\mathbb{E}(D)$ si bien que $\mathbb{V}\mathrm{ar}(\mathbb{E}(R \mid N)) = \mathbb{V}\mathrm{ar}(N)(\mathbb{E}(D))^2$ et

$$\mathbb{V}\mathrm{ar}(R \mid N) = \mathbb{E}\left[\left(R - N\mathbb{E}(D)\right)^2 \mid N\right] = \mathbb{E}\left[\left(\sum_{k=1}^{N}(D_i - \mathbb{E}(D))\right)^2 \mid N\right] = N\mathbb{V}\mathrm{ar}(D)$$

et donc

$$\mathbb{V}\mathrm{ar}(R) = \mathbb{V}\mathrm{ar}(N)(\mathbb{E}(D))^2 + \mathbb{E}(N)\mathbb{V}\mathrm{ar}(D) = 400 \times 10^4 + 400 \times 10^4 = 8 \times 10^5.$$

Exercice 1.4

1. Par définition,

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

et donc

$$\phi'_X(x) = \sum_{k \geq 0} k z^{k-1} \mathbb{P}(X = k) = \mathbb{E}\left(X z^{X-1}\right) \quad \text{et} \quad \phi''_X(x) = \sum_{k \geq 0} k(k-1) z^{k-2} \mathbb{P}(X = k) = \mathbb{E}\left(X(X-1) z^{X-2}\right)$$

ce qui donne

$$\phi'_X(1) = \mathbb{E}\left(X\right) \quad \text{et} \quad \phi''_X(x) = \mathbb{E}\left(X(X-1)\right).$$

Exercice 1.5

1. Si G_k est le nombre de personnes impliquées dans le k-ième accident, alors $Z_t = \sum_{k=1}^{X_t} G_k$ et par hypothèse, les G_k sont i.i.d. et indépendantes de X_t. Ainsi, la question 6 de l'exercice 1 donne

$$\mathbb{E}\left(z^{Z_t} \mid X_t\right) = \mathbb{E}\left(\prod_{k=1}^{X_t} z^{G_1} \mid X_t\right) = \mathbb{E}\left(z^{G_1}\right)^{X_t} = \phi_{G_1}(z)^{X_t}$$

puis le théorème de l'espérance totale donne

$$\phi_{Z_t}(z) = \mathbb{E}\left[\mathbb{E}(z^{Z_t} \mid X_t)\right] = \mathbb{E}\left[\phi_{G_1}(z)^{X_t}\right] = \phi_{X_t}\left(\phi_{G_1}(z)\right).$$

2. Puisque $\phi_{X_t}(z) = e^{-at(1-z)}$ et $\phi_{G_1}(z) = z/(2-z)$, il vient

$$\phi'_{Z_t} = at\phi'_{G_1}e^{-at(1-\phi_{G_1})} \text{ et } \phi''_{Z_t} = at\phi''_{G_1}e^{-a(1-\phi_{G_1})} + a^2t^2(\phi'_{G_1})^2e^{-at(1-\phi_{G_1})}.$$

Puisque

$$\phi'_{G_1}(z) = \frac{2}{(2-z)^2} \text{ et } \phi''_{G_1}(z) = \frac{4}{(2-z)^3},$$

on obtient finalement $\mathbb{E}(Z_t) = \phi'_{Z_t}(1) = 2at$ et

$$\mathbb{V}\mathrm{ar}(Z_t) = \phi''_{Z_t}(1) + \phi'_{Z_t}(1)(1-\phi'_{Z_t}(1)) = 4at + 4a^2t^2 + 2at(1-2at) = 6at.$$

3. On a

$$\mathbb{E}(Z_t) = \mathbb{E}(X_t)\mathbb{E}(G) = 2at$$

et

$$\mathbb{V}\mathrm{ar}(Z_t) = \mathbb{E}(X_t\mathbb{V}\mathrm{ar}(G)) + \mathbb{V}\mathrm{ar}(X_t\mathbb{E}(G)) = 2at + 4at = 6at.$$

Problème 1.6

1. Conditionnellement à $X = k$, Y suit une loi binomiale de paramètre $(20-k, 1/4)$ puisque pour chacune des $20-k$ questions que le candidat n'a pas apprise, la probabilité d'obtenir un point vaut $1/4$. On a donc

$$\mathbb{E}(Y \mid X) = 5 - \frac{X}{4} \text{ et } \mathbb{V}\mathrm{ar}(Y \mid X) = \frac{3(20-X)}{16}.$$

2. On a

$$\mathbb{E}(N) = \mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(\mathbb{E}(Y \mid X)) = 5 + \frac{3\mathbb{E}(X)}{4}$$

et

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(N) &= \mathbb{E}(\mathbb{V}\mathrm{ar}(N \mid X)) + \mathbb{V}\mathrm{ar}(\mathbb{E}(N \mid X)) \\
&= \mathbb{E}(\mathbb{V}\mathrm{ar}(X+Y \mid X)) + \mathbb{V}\mathrm{ar}(\mathbb{E}(X+Y \mid X)) \\
&= \mathbb{E}(\mathbb{V}\mathrm{ar}(Y \mid X)) + \mathbb{V}\mathrm{ar}(X + \mathbb{E}(Y \mid X)) \\
&= \frac{3(20-\mathbb{E}(X))}{16} + \mathbb{V}\mathrm{ar}\left(5 + \frac{3X}{4}\right) \\
&= \frac{3(20-\mathbb{E}(X))}{16} + \frac{9}{16}\mathbb{V}\mathrm{ar}\left(X\right).
\end{aligned}$$

3. On a $\binom{100}{20}$ questionnaires possibles. Si $X = k$, cela signifie que parmi les 20 questions, on en a tirées k parmi les p connues du candidat et $20-k$ parmi les $100-p$ autres ce qui donne le résultat attendu. Pour la moyenne, cela donne

$$\mathbb{E}(X) = \frac{1}{\binom{100}{20}}\sum_{k=0}^{20} k\binom{p}{k}\binom{100-p}{20-k}.$$

Pour se ramener à l'identité de l'indication, il faut faire disparaître le k : pour cela on utilise $k\binom{p}{k} = p\binom{p-1}{k-1}$ et l'on obtient

$$\mathbb{E}(X) = \frac{p}{\binom{100}{20}}\sum_{k=0}^{20}\binom{p-1}{k-1}\binom{100-p}{20-k} = \frac{p}{\binom{100}{20}}\sum_{k=0}^{20}\binom{p-1}{k-1}\binom{100-p}{19-(k-1)} = \frac{p}{\binom{100}{20}}\binom{99}{19} = \frac{p}{5}.$$

Pour la variance, on calcule

$$\mathbb{E}(X(X-1)) = \frac{1}{\binom{100}{20}}\sum_{k=0}^{20} k(k-1)\binom{p}{k}\binom{100-p}{20-k} = \frac{p(p-1)}{\binom{100}{20}}\sum_{k=0}^{20}\binom{p-2}{k-2}\binom{100-p}{20-k} = \frac{p(p-1)}{\binom{100}{20}}\binom{98}{18}$$

et donc $\mathbb{E}(X(X-1)) = 19p(p-1)/495$. On en déduit

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}(X(X-1)) + \mathbb{E}(X) - (\mathbb{E}(X))^2 = \frac{19p(p-1)}{495} + \frac{p}{5} - \frac{p^2}{25} = \frac{p(100-p)}{725}.$$

4. Pour $p = 50$ on a alors $\mathbb{E}(N) = 12{,}5$ et $\mathbb{V}\mathrm{ar}(N) \approx 4{,}15$, et pour $p = 70$ on a $\mathbb{E}(N) = 15{,}5$ et $\mathbb{V}\mathrm{ar}(N) \approx 3{,}03$. N'apprendre que 50 questions n'assure pas d'avoir la moyenne !

A.2 Exercices du Chapitre 2

↪ **Exercice 2.1**

1. On a

$$\mathbb{P}(f_X(X) = 0) = \int \mathbb{1}\{f_X(x) = 0\} f_X(x)\mathrm{d}x$$

qui vaut bien 0 car $\mathbb{1}\{f_X(x) = 0\} f_X(x) = 0$. Ce résultat garantit que la densité conditionnelle $f_{Y|X}$ de Y sachant X est bien définie.

↪ **Exercice 2.2** *(Calcul de densités)*

1. Soit $g(x) = \frac{1}{|a|} f_X\left(\frac{x-b}{a}\right)$: le théorème de transfert (Théorème 2.6.3) donne

$$\mathbb{P}(aX + b \leq z) = \int \mathbb{1}\{ax + b \leq z\} f_X(x)\mathrm{d}x$$

et donc par changement de variable, (attention à la valeur absolue ! !)

$$\mathbb{P}(aX + b \leq z) = \frac{1}{|a|} \int \mathbb{1}\{y \leq z\} f_X\left(\frac{y-b}{a}\right) \mathrm{d}x = \int_{-\infty}^{z} g.$$

Puisque $F_{aX+b}(z) = \int_{-\infty}^{z} f_{aX+b}$, on obtient en dérivant $f_{aX+b} = g$ ce qui donne le résultat.

2. Soit $g(x) = (f_X(x) + f_X(-x))\mathbb{1}\{x \geq 0\}$: comme pour la question précédente il suffit de montrer que $F_{|X|}(x) = \int_{-\infty}^{x} g$. Pour $x < 0$ on a $F_{|X|}(x) = \mathbb{P}(|X| \leq x) = 0$ et pour $x \geq 0$,

$$F_{|X|}(x) = \mathbb{P}(|X| \leq x) = \mathbb{P}(-x \leq X \leq x) = \int_{-x}^{x} f_X(a)\mathrm{d}a = \int_{0}^{x} (f_X(a) + f_X(-a))\mathrm{d}a$$

ce qui montre que $F_X(x) = \int_{-\infty}^{x} g$ pour tout $x \in \mathbb{R}$.

3. Soit $g(s) = \int f_X(x) f_Y(s-x)\mathrm{d}x$: comme précédemment il suffit de montrer que $F_{X+Y}(z) = \int_{-\infty}^{z} g$. Puisque X et Y sont indépendantes et que chacune est absolument continue, le couple (X, Y) est absolument continue (Théorème 2.6.5). Par définition de la densité (en dimension 2) ou bien en utilisant le théorème de transfert (Théorème 2.6.3) avec $\varphi(x, y) = \mathbb{1}\{x + y \leq z\}$, on obtient

$$\mathbb{P}(X + Y \leq z) = \int \mathbb{1}\{x + y \leq z\} f_{X,Y}(x, y)\mathrm{d}x\mathrm{d}y.$$

Puisque X et Y sont indépendantes, la densité du couple est égale au produit des densités (Théorème 2.6.5) et donc

$$\mathbb{P}(X + Y \leq z) = \int \mathbb{1}\{x + y \leq z\} f_X(x) f_Y(y)\mathrm{d}x\mathrm{d}y.$$

Le changement de variables $(x, y) \mapsto (x, x + y)$ donne

$$\mathbb{P}(X + Y \leq z) = \int \mathbb{1}\{s \leq z\} f_X(x) f_Y(s - x)\mathrm{d}x\mathrm{d}s$$

et puisque toutes les quantités sont ≥ 0, on peut utiliser le théorème de Fubini qui donne

$$\mathbb{P}(X + Y \leq z) = \int \mathrm{d}s \mathbb{1}\{s \leq z\} g(s)\mathrm{d}s$$

ce qui prouve le résultat.

4. On a $\mathbb{P}(\max(X,Y) \leq z) = \mathbb{P}(X \leq z)\mathbb{P}(Y \leq z)$ et il suffit de dériver : la dérivée du membre de gauche donne la densité de $\max(X,Y)$ et la densité du membre de droite est égale à $f_X(z)\mathbb{P}(Y \leq z) + f_Y(z)\mathbb{P}(X \leq z)$ ce qui prouve le résultat.

5. Pour $x < 0$ on a $\mathbb{P}(-c\ln(U) \leq x) = 0$ et pour $x \geq 0$,

$$\mathbb{P}(-c\ln(U) \leq x) = \mathbb{P}(U \geq e^{-x/c}) = 1 - e^{-x/c}.$$

La dérivée vaut $(1/c)e^{-x/c}$: on reconnaît la densité de la loi exponentielle de paramètre $1/c$.

6. L'application

$$\varphi : (u_1, u_2) \in]0,1[^2 \mapsto \left(\cos(2\pi u_1)\sqrt{-2\ln(u_2)}, \sin(2\pi u_1)\sqrt{-2\ln(u_2)}\right) \in \Delta$$

avec $\Delta = \mathbb{R}^2 \setminus \{(x,0) : x \geq 0\}$ est un difféomorphisme C^∞ d'inverse

$$\varphi^{-1} : x = (x_1, x_2) \in \Delta \mapsto \left(\frac{1}{2\pi}\theta(x), e^{-x^T x/2}\right) \in]0,1[^2$$

avec $\theta(x) \in]0, 2\pi[$ tel que $\cos\theta(x) = x_1/(x^T x)^{1/2}$ et $\sin\theta(x) = x_2/(x^T x)^{1/2}$, en particulier

$$\theta(x) = \arctan\left(\frac{x_2}{x_1}\right) + \begin{cases} 0 & \text{si } x_1, x_2 > 0, \\ 2\pi & \text{si } x_1 > 0 > x_2, \\ \pi & \text{si } x_2 > 0 > x_1. \end{cases}$$

En particulier, θ a les mêmes dérivées partielles que $(x_1, x_2) \mapsto \arctan(x_2/x_1)$ et on en déduit que

$$\mathrm{Jac}_x\left(\varphi^{-1}\right) = \begin{pmatrix} -\dfrac{x_2}{2\pi x^T x} & \dfrac{x_1}{2\pi x^T x} \\ -x_1 e^{-x^T x/2} & -x_2 e^{-x^T x/2} \end{pmatrix}$$

et donc

$$\det\left(\mathrm{Jac}_x\left(\varphi^{-1}\right)\right) = \frac{x_2^2}{2\pi x^T x}e^{-x^T x} + \frac{x_1^2}{2\pi x^T x}e^{-x^T x} = \frac{1}{2\pi}e^{-x^T x/2}$$

et le Théorème 2.5.1 donne donc

$$f_{X_1,X_2}(x) = \frac{1}{2\pi}e^{-x^T x} \quad \text{avec} \quad (X_1, X_2) = \varphi(U_1, U_2).$$

Puisque

$$f_{X_1,X_2}(x_1, x_2) = \left(\frac{1}{\sqrt{2\pi}}e^{-x_1^2/2}\right)\left(\frac{1}{\sqrt{2\pi}}e^{-x_2^2/2}\right)$$

on voit que X_1 et X_2 sont i.i.d. et suivent la loi normale standard d'après le Théorème 2.6.5.

↪ **Exercice 2.3** *(Loi normale)*

1. De manière générale, on a par le théorème de transfert

$$\mathbb{E}(X) = \int x f_X(x)\mathrm{d}x \quad \text{et} \quad \mathbb{V}\mathrm{ar}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \int (x - \mathbb{E}(X))^2 f_X(x)\mathrm{d}x.$$

Dans le cas présent, cela donne donc

$$\mathbb{E}(X) = \int x \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)\mathrm{d}x$$

et le changement de variables $y = (x-m)/\sigma$ donne donc

$$\mathbb{E}(X) = \int (\sigma y + m)\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)\mathrm{d}y = \sigma \int y \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)\mathrm{d}y + m \int \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)\mathrm{d}y.$$

La première intégrale est nulle par parité, et la seconde vaut 1 puisque $\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)$ est une densité (c'est celle de la loi normale standard avec $m=0$ et $\sigma=1$). Cela montre bien que $\mathbb{E}(X)=m$. Pour la variance, le même changement de variable donne, et en utilisant aussi $m=\mathbb{E}(X)$,

$$\mathbb{V}\mathrm{ar}(X)=\mathbb{E}((X-\mathbb{E}(X))^2)=\mathbb{E}((X-m)^2)=\sigma^2\int y^2\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)\mathrm{d}y$$

et par intégration par parties,

$$\mathbb{V}\mathrm{ar}(X)=\mathbb{E}((X-\mathbb{E}(X))^2)=\sigma^2\left[-ye^{-y^2/2}\right]_{-\infty}^{+\infty}+\sigma^2\int\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)\mathrm{d}y$$

ce qui donne bien $\mathbb{V}\mathrm{ar}(X)=\sigma^2$.

2. La formule de transfert donne

$$\mathbb{E}(X^k)=\int x^k f_X(x)\mathrm{d}x.$$

Lorsque $m=0$, f_X est paire et donc si k est impair, $x^k f_X(x)$ est impaire et son intégrale est donc nulle.

3. Pour $k=1$ on a $\mathbb{E}(X^2)=\mathbb{V}\mathrm{ar}(X)+(\mathbb{E}(X))^2=\sigma^2+m^2$ qui vaut bien 1 puisque $\sigma=1$ et $m=0$. Pour k plus grand, le théorème de transfert donne

$$\mathbb{E}(X^{2k+2})=\frac{1}{\sqrt{2\pi}}\int x^{2k+2}e^{-x^2/2}\mathrm{d}x$$

et on obtient donc en intégrant par parties

$$\mathbb{E}(X^{2k+2})=\frac{1}{\sqrt{2\pi}}\left(\left[-x^{2k+1}e^{-x^2/2}\right]_{-\infty}^{\infty}+(2k+2)\int x^{2k}e^{-x^2/2}\mathrm{d}x\right)$$

ce qui donne bien $\mathbb{E}(X^{2k+2})=(2k+1)\mathbb{E}(X^{2k})$.

4. Le Théorème 2.7.2 donne

$$\mathbb{E}(X\mid a<X<b)=\frac{\mathbb{E}(X\mathbb{1}\{a<X<b\})}{\mathbb{P}(a<X<b)}.$$

Puisque $\mathbb{P}(a<X<b)=F_X(b)-F_X(a)$, il suffit de montrer que le numérateur est égal à $f_X(a)-f_X(b)$. Pour cela, on utilise le théorème de transfert qui nous donne

$$\mathbb{E}(X\mathbb{1}\{a<X<b\})=\frac{1}{\sqrt{2\pi}}\int_a^b xe^{-x^2/2}\mathrm{d}x=\frac{1}{\sqrt{2\pi}}\left[-e^{-x^2/2}\right]_a^b$$

qui vaut bien $f_X(a)-f_X(b)$.

5. On regarde les fonctions caractéristiques : par définition,

$$\varphi_{X+Y}(t)=\mathbb{E}(e^{it(X+Y)})=\mathbb{E}(e^{itX}e^{itY}).$$

Puisque X et Y sont indépendantes, e^{itX} et e^{itY} le sont aussi (Proposition 1.4.3) et donc l'espérance du produit est égal au produit des espérances (Proposition 1.4.4) : cela donne

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

En utilisant l'expression de la fonction caractéristique d'une loi normale (Proposition 2.5.6) on obtient donc

$$\varphi_{X+Y}(t)=\exp\left(-\frac{-t^2}{2}(\mathbb{V}\mathrm{ar}(X)+\mathbb{V}\mathrm{ar}(Y))+ix(\mathbb{E}(X)+\mathbb{E}(Y))\right).$$

On reconnaît la fonction caractéristique d'une loi normale de moyenne $\mathbb{E}(X)+\mathbb{E}(Y)=\mathbb{E}(X+Y)$ et de variance $\mathbb{V}\mathrm{ar}(X)+\mathbb{V}\mathrm{ar}(Y)=\mathbb{V}\mathrm{ar}(X+Y)$ (puisque X et Y sont indépendantes).

↪ **Exercice 2.4** *(Lois Gamma et Beta)*

1. Pour que les fonctions considérées soient des densités, il faut qu'elles soient ≥ 0 et que leur intégrale vaille 1. Par définition, les fonctions $f^{\Gamma}_{\alpha,\lambda}$ et $f^{\beta}_{\alpha,\beta}$ sont ≥ 0, et pour qu'elles soient intégrables il faut $\alpha > 0$ et $\lambda \geq 0$ pour la loi Gamma, et $\alpha, \beta > 0$ pour la loi beta. Sous ces conditions, les valeurs ci-dessus pour $c^{\Gamma}_{\alpha,\lambda}$ et $c^{\beta}_{\alpha,\beta}$ sont les constantes de normalisation qui assurent que

$$\int f^{\Gamma}_{\alpha,\lambda}(x)\mathrm{d}x = \int f^{\beta}_{\alpha,\beta}(x)\mathrm{d}x = 1.$$

2. La première identité s'obtient par intégration par partie, et pour la deuxième identité :

$$\begin{aligned}\Gamma(\alpha)\Gamma(\beta) &= \int x^{\alpha-1}y^{\alpha-1}e^{-(x+y)}\mathbb{1}\left\{x, y \geq 0\right\}\mathrm{d}x\mathrm{d}y \\ &= \int (zt)^{\alpha-1}(z(1-t))^{\beta-1}e^{-z}z\mathbb{1}\left\{z \geq 0, t \in [0,1]\right\}\mathrm{d}z\mathrm{d}t\end{aligned}$$

où l'on a fait le changement de variable $(x, y) = \varphi^{-1}(z, t)$ et le facteur z correspond à la valeur absolue du déterminant du Jacobien de φ^{-1}.

3. Loi Gamma : la moyenne est donnée par

$$\mathbb{E}(X) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)}\int_0^{\infty} x^{\alpha}e^{-\lambda x}\mathrm{d}x = \frac{1}{\lambda\Gamma(\alpha)}\int_0^{\infty} x^{\alpha}e^{-x}\mathrm{d}x = \frac{\alpha}{\lambda}$$

et

$$\mathbb{E}(X^2) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)}\int_0^{\infty} x^{\alpha+1}e^{-\lambda x}\mathrm{d}x = \frac{1}{\lambda^2\Gamma(\alpha)}\int_0^{\infty} x^{\alpha+1}e^{-x}\mathrm{d}x = \frac{\alpha(\alpha+1)}{\lambda^2}$$

si bien que la variance est donnée par

$$\mathbb{V}\mathrm{ar}(X) = \frac{\alpha^2+\alpha}{\lambda^2} - \frac{\alpha^2}{\lambda^2} = \frac{\alpha}{\lambda^2}.$$

La transformée de Laplace :

$$\mathbb{E}(e^{-sX}) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)}\int_0^{\infty} e^{-sx}x^{\alpha-1}e^{-\lambda x}\mathrm{d}x = \frac{\lambda^{\alpha}}{(\lambda+s)^{\alpha}}$$

Loi Beta : la moyenne est donnée par

$$\mathbb{E}(X) = \frac{1}{B(\alpha,\beta)}\int_0^1 x^{\alpha}(1-x)^{\beta-1}\mathrm{d}x = \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} = \frac{\alpha}{\alpha+\beta}$$

et

$$\mathbb{E}(X^2) = \frac{1}{B(\alpha,\beta)}\int_0^1 x^{\alpha+1}(1-x)^{\beta-1}\mathrm{d}x = \frac{\Gamma(\alpha+2)\Gamma(\beta)}{\Gamma(\alpha+\beta+2)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} = \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)}$$

et donc la variance est donnée par

$$\begin{aligned}\mathbb{V}\mathrm{ar}(X) &= \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)} - \frac{\alpha^2}{(\alpha+\beta)^2} \\ &= \frac{1}{(\alpha+\beta+1)(\alpha+\beta)^2}\left((\alpha^2+\alpha)(\alpha+\beta) - \alpha^3 - \alpha^2\beta - \alpha^2\right) \\ &= \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}.\end{aligned}$$

Pour la transformée de Laplace,

$$\begin{aligned}\mathbb{E}(e^{-sX}) &= \frac{1}{B(\alpha,\beta)}\int_0^1 e^{-sx}x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x \\ &= \sum_{k\geq 0}\frac{(-s)^k}{k!}\frac{B(\alpha+k,\beta)}{B(\alpha,\beta)} \\ &= \sum_{k\geq 0}\frac{(-s)^k}{k!}\frac{\Gamma(\alpha+k)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha+\beta+k)} \\ &= 1 + \sum_{k\geq 1}\frac{(-s)^k}{k!}\prod_{i=0}^{k-1}\frac{\alpha+i}{\alpha+\beta+i}.\end{aligned}$$

4. On a

$$\begin{aligned}\mathbb{E}\left(f\left(\frac{U}{U+V}\right)\right) &= \frac{1}{\Gamma(\alpha+\beta)B(\alpha,\beta)}\int f\left(\frac{u}{u+v}\right)u^{\alpha-1}v^{\beta-1}e^{-(u+v)}\mathbb{1}\left\{u,v\geq 0\right\}\mathrm{d}u\mathrm{d}v\\ &= \frac{1}{\Gamma(\alpha+\beta)B(\alpha,\beta)}\int f(t)(zt)^{\alpha-1}(z(1-t))^{\beta-1}e^{-z}z\mathbb{1}\left\{z\geq 0, t\in[0,1]\right\}\mathrm{d}z\mathrm{d}t\\ &= \frac{1}{\Gamma(\alpha+\beta)B(\alpha,\beta)}\int_0^1 f(t)t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t\int_0^\infty e^{-z}z^{\alpha+\beta-1}\mathrm{d}z\\ &= \frac{1}{B(\alpha,\beta)}\int_0^1 f(t)t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t.\end{aligned}$$

En prenant f la fonction indicatrice d'un Borélien on obtient donc la réponse.

↪ **Exercice 2.5** *(Loi exponentielle)*

1. On a

$$\mathbb{E}\left(\exp\left(-\lambda\sum_{k=1}^n E_k\right)\right) = \left(\mathbb{E}(e^{-\lambda E_1})\right)^n = \left(\frac{\mu}{\lambda+\mu}\right)^n$$

et l'on reconnaît donc la transformée de Laplace d'une loi Gamma de paramètre (n,μ).

2. Si l'on définit $q=\mu/(\lambda+\mu)$ et que l'on note p le paramètre de G,

$$\mathbb{E}\left(\exp\left(-\lambda\sum_{k=1}^G E_k\right)\right) = \mathbb{E}\left(\left(\frac{\mu}{\lambda+\mu}\right)^G\right) = \sum_{k\geq 1}(1-p)^{k-1}pq^k = \frac{pq}{1-q(1-p)}$$

ce qui donne

$$\mathbb{E}\left(\exp\left(-\lambda\sum_{k=1}^G E_k\right)\right) = \frac{p\mu}{\lambda+\mu-\mu(1-p)} = \frac{p\mu}{\lambda+\mu p}.$$

On reconnaît la transformée de Laplace de la loi exponentielle de paramètre μp : ce n'est rien d'autre que la propriété de division du processus de Poisson !

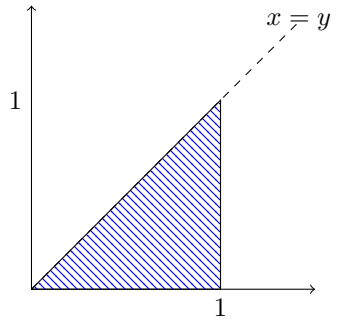
3. Pour $x\in]0,1[$, on a

$$\mathbb{P}\left(\frac{E_1}{E_1+E_2}\geq x\right) = \mathbb{P}(E_2\leq E_1(1/x-1)) = \mathbb{E}\left(1-e^{-\mu(1/x-1)E_1}\right) = 1-\frac{\mu}{\mu+\mu(1/x-1)} = 1-x.$$

On reconnaît donc une loi uniforme sur $[0,1]$.

↪ **Exercice 2.6**

1. Δ est la région hachurée sur le dessin suivant :



f est ≥ 0, le changement de variables $(x,y)\in\Delta\mapsto(x,y/x)\in]0,1]^2$ donne

$$\int f = \int\sqrt{\frac{x}{y}}\mathbb{1}\left\{0<y\leq x\leq 1\right\}\mathrm{d}x\mathrm{d}y = \int\frac{x}{\sqrt{z}}\mathbb{1}\left\{0<x,z\leq 1\right\}\mathrm{d}x\mathrm{d}z = \left(\int_0^1 x\mathrm{d}x\right)\left(\int_0^1\frac{\mathrm{d}z}{\sqrt{z}}\right) = 1$$

et donc f est bien une densité. Le Théorème 2.6.4 donne pour $x \in]0,1]$

$$f_X(x) = \int f_{X,Y}(x,y)\mathrm{d}y = \int \sqrt{\frac{x}{y}} \mathbb{1}\{0 < y \leq x\}\,\mathrm{d}y = \sqrt{x}\,[2\sqrt{y}]_0^x = 2x$$

et pour $y \in]0,1]$

$$f_Y(y) = \int f_{X,Y}(x,y)\mathrm{d}x = \int \sqrt{\frac{x}{y}} \mathbb{1}\{y \leq x \leq 1\}\,\mathrm{d}x = \frac{1}{\sqrt{y}}\left[\frac{2}{3}x^{3/2}\right]_y^1 = \frac{2(1-y^{3/2})}{3y^{1/2}}.$$

2. On a $f_{X,Y}(x,y) \neq f_X(x)f_Y(y)$ et donc X et Y ne sont pas indépendantes au vu du Théorème 2.6.5. On a

$$\mathbb{E}(X) = \int_0^1 x f_X(x)\mathrm{d}x = 2\int_0^1 x^2\mathrm{d}x = \frac{2}{3}$$

et

$$\mathbb{E}(Y) = \int_0^1 y f_Y(y)\mathrm{d}y = \frac{2}{3}\int_0^1 y^{1/2}(1-y^{3/2})\mathrm{d}y = \frac{4}{9}\int_0^1 (1-y^{3/2})\mathrm{d}(y^{3/2}) = \frac{4}{9}\int_0^1 (1-z)\mathrm{d}z = \frac{2}{9}.$$

Par ailleurs, le théorème de transfert 2.6.3 donne

$$\begin{aligned}
\mathbb{E}(XY) &= \int xy f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y \\
&= \int x^{3/2}y^{1/2}\mathbb{1}\{0 < y \leq x \leq 1\}\,\mathrm{d}x\mathrm{d}y \\
&= \int_0^1 x^{3/2}\left[\frac{2}{3}y^{3/2}\right]_0^x \mathrm{d}x \\
&= \frac{2}{3}\int_0^1 x^3\mathrm{d}x
\end{aligned}$$

et donc $\mathbb{E}(XY) = 1/6$. Finalement,

$$\mathbb{C}\mathrm{ov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \frac{1}{6} - \frac{4}{27} = \frac{1}{54}.$$

3. Soit $\varphi : (x,y) \in \Delta \mapsto (x, y/x) \in]0,1]^2$ d'inverse $\varphi^{-1} : (x,z) \in]0,1]^2 \mapsto (x, xz) \in \Delta$: alors

$$\mathrm{Jac}_{x,z}\left(\varphi^{-1}\right) = \begin{pmatrix} 1 & 0 \\ z & x \end{pmatrix}$$

dont le déterminant vaut x, si bien que par le Théorème 2.5.1, on a pour $x, z \in]0,1[$

$$f_{X,Y/X}(x,z) = f_{X,Y}(x,xz)x = x\sqrt{\frac{x}{xz}} = \frac{x}{\sqrt{z}}.$$

En intégrant sur x, le Théorème 2.6.4 donne

$$f_{Y/X}(z) = \int_0^1 \frac{x}{\sqrt{z}}\mathrm{d}x = \frac{1}{2\sqrt{z}}.$$

On a bien $f_{X,Y/X} = f_X f_{Y/X}$ et donc X et Y/X sont indépendantes.

4. Par définition, la densité $f_{Y|X}$ de la loi de Y sachant X est donnée par

$$f_{Y|X}(y) = \frac{f_{X,Y}(X,y)}{f_X(X)} = \sqrt{\frac{X}{y}}\frac{1}{2X}\mathbb{1}\{0 < y \leq X\} = \frac{1}{2\sqrt{yX}}\mathbb{1}\{0 < y \leq X\}$$

et donc le Théorème 2.7.4 donne

$$\mathbb{E}(Y \mid X) = \int_0^X \frac{y}{2\sqrt{yX}}\mathrm{d}y = \frac{1}{2\sqrt{X}}\left[\frac{2}{3}y^{3/2}\right]_0^X = \frac{X}{3}.$$

A l'aide du théorème de l'espérance totale, on retrouve

$$\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y \mid X)) = \frac{1}{3}\mathbb{E}(X) = \frac{1}{3}\frac{2}{3} = \frac{2}{9}.$$

5. Par définition de la densité,

$$\begin{aligned}\mathbb{P}(Y < X/2) &= \int \mathbb{1}\{y < x/2\}\sqrt{\frac{x}{y}}\mathbb{1}\{0 < y \leq x \leq 1\}\,\mathrm{d}x\mathrm{d}y \\ &= \int \frac{x}{\sqrt{z}}\mathbb{1}\{0 < z \leq 1/2, 0 < x \leq 1\}\,\mathrm{d}x\mathrm{d}z \\ &= \frac{1}{\sqrt{2}}.\end{aligned}$$

Par définition, la densité de X sachant Y est donnée par

$$f_{X|Y}(x) = \frac{f_{X,Y}(x,Y)}{f_Y(Y)} = \sqrt{\frac{x}{Y}}\frac{3Y^{1/2}}{2(1-Y^{3/2})} = \frac{3\sqrt{x}}{2(1-Y^{3/2})}.$$

6. On en déduit comme précédemment

$$\mathbb{E}(X \mid Y) = \int_Y^1 x\frac{3\sqrt{x}}{2(1-Y^{3/2})}\mathrm{d}x = \frac{3(1-Y^{5/2})}{5(1-Y^{3/2})}$$

↪ Exercice 2.7

1. On a

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

et donc par hypothèse,

$$f_{X,Y}(x,y) = f_{Y|X=x}(y)f_X(x) = \frac{1}{\sqrt{2\pi\frac{1}{2x}}}e^{-y^2/(2\frac{1}{2x})} \times \frac{1}{\sqrt{\pi x}}e^{-x}\mathbb{1}\{x > 0\} = \frac{1}{\pi}e^{-x(1+y^2)}\mathbb{1}\{x > 0\}$$

d'où on déduit

$$f_Y(y) = \int_0^\infty \frac{1}{\pi}e^{-x(1+y^2)}\mathrm{d}x = \frac{1}{\pi(1+y^2)}.$$

2. On a

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{1}{\pi}e^{-x(1+y^2)}\mathbb{1}\{x > 0\} \times \pi(1+y^2) = (1+y^2)e^{-x(1+y^2)}\mathbb{1}\{x > 0\}$$

d'où on déduit

$$f_{X|Y}(x) = (1+Y^2)e^{-x(1+Y^2)}\mathbb{1}\{x > 0\}$$

et donc

$$\mathbb{E}(X \mid Y) = \int_0^\infty x(1+Y^2)e^{-x(1+Y^2)}\mathrm{d}x = \frac{1}{1+Y^2}\int_0^\infty xe^{-x}\mathrm{d}x = \frac{1}{1+Y^2}.$$

Par le théorème de l'espérance totale,

$$\mathbb{E}(X) = \mathbb{E}\left(\frac{1}{1+Y^2}\right)$$

et par le théorème de transfert,

$$\mathbb{E}(X) = \int \frac{1}{1+y^2}f_Y(y)\mathrm{d}y = \frac{1}{\pi}\int \frac{\mathrm{d}y}{(1+y^2)^2}.$$

D'un autre côté, si on repart de la densité de X on obtient

$$\mathbb{E}(X) = \frac{1}{\sqrt{\pi}}\int_0^\infty x^{1/2}e^{-x}\mathrm{d}x = \frac{1}{\sqrt{\pi}}\Gamma(3/2).$$

On obtient donc

$$\int \frac{\mathrm{d}y}{(1+y^2)^2} = \sqrt{\pi}\Gamma(3/2).$$

↪ **Exercice 2.8**

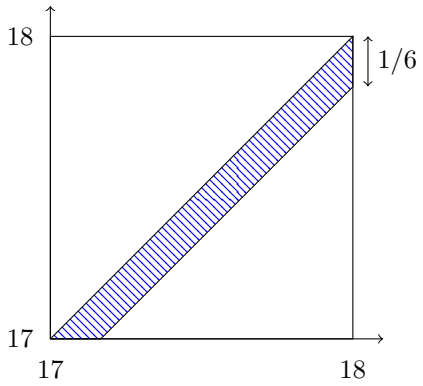
1. On note T_1 et T_2 les instants d'arrivée : on cherche donc à calculer

$$\mathbb{P}(\text{rencontre}) = \mathbb{P}(\max(T_1, T_2) \leq \min(T_1, T_2) + 1/6).$$

Par symétrie,

$$\begin{aligned}\mathbb{P}(\max(T_1, T_2) \leq \min(T_1, T_2) + 10) &= 2\mathbb{P}(T_2 \leq T_1 + 1/6, T_1 \leq T_2) \\ &= 2\int \mathbb{1}\left\{17 \leq t_2 \leq t_1 \leq 18, t_2 \geq t_1 - 1/6\right\} \mathrm{d}t_1 \mathrm{d}t_2.\end{aligned}$$

On cherche donc à calculer l'aire ci-dessous :



qui vaut $y\sqrt{2} - y^2$ avec $2y^2 = 1/6^2$, soit $11/72$. La réponse est donc $\mathbb{P}(\text{rencontre}) = 11/36$.

2. En généralisant le raisonnement ci-dessus, on obtient pour $x \in [0, 1]$

$$\mathbb{P}(|T_1 - T_2| \leq x) = 2\left(y\sqrt{2} - y^2\right)$$

avec y donné par $2y^2 = x^2$, soit

$$\mathbb{P}(|T_1 - T_2| \leq x) = 2x - x^2.$$

La densité de $|T_1 - T_2|$ vaut donc $2(1-x)\mathbb{1}\{x \in [0,1]\}$.

Problème 2.9

1. Soit $x \geq 0$: on a

$$\mathbb{P}(E_1 - x \geq y \mid E_1 \geq x) = \frac{\mathbb{P}(E_1 \geq x + y)}{\mathbb{P}(E_1 \geq x)} = e^{-\mu y}.$$

2. On a $\mathbb{P}(E_1/\alpha \geq x) = \mathbb{P}(E_1 \geq \alpha x) = e^{-\mu\alpha x}$.

3. On a

$$\begin{aligned}\mathbb{E}(e^{-\lambda E_2 - \lambda'(E_1 - E_2)} \mid E_2 \leq E_1) &= 2\int \mathbb{1}\left\{0 \leq x_2 \leq x_1\right\} e^{-\lambda x_2 - \lambda' x_1 - x_2} \mu^2 e^{-\mu(x_1+x_2)} \mathrm{d}x_1 \mathrm{d}x_2 \\ &= 2\int \mathbb{1}\left\{0 \leq u, v\right\} e^{-\lambda u - \lambda' v} \mu^2 e^{-\mu(v+2u)} \mathrm{d}u\mathrm{d}v \\ &= \left(\int 2\mu e^{-2\mu u} e^{-\lambda u} \mathrm{d}u\right)\left(\int \mu e^{-\mu v} e^{-\lambda' v} \mathrm{d}v\right) \\ &= \mathbb{E}(e^{-\lambda E_1/2})\mathbb{E}(e^{-\lambda' E_1}).\end{aligned}$$

4. Cela montre bien que $(E_2, E_1 - E_2)$ conditionnellement à $E_2 \leq E_1$ est égale en distribution à $(E_2/2, E_1)$. Pour le max, on écrit

$$\begin{aligned}\mathbb{P}(\max(E_1, E_2) \geq x) &= 2\mathbb{P}(E_1 \geq x, E_2 \leq E_1) \\ &= \mathbb{P}(E_1 \geq x \mid E_2 \leq E_1) \\ &= \mathbb{P}(E_2 + (E_1 - E_2) \geq x \mid E_2 \leq E_1) \\ &= \mathbb{P}(E_1 + E_2/2 \geq x)\end{aligned}$$

où l'on a utilisé la question précédente pour la troisième égalité.

5. On a

$$\begin{aligned}&\mathbb{E}(e^{-\lambda_n E_n - \lambda_1(E_1 - E_n) - \cdots - \lambda_{n-1}(E_{n-1} - E_n)} \mid E_n \leq E_k, k = 1, \ldots, n) \\ &\quad = n \int \mathbb{1}\{0 \leq x_n \leq x_k, k = 1, \ldots, n\} e^{-\lambda_n x_n - \lambda_1(x_1 - x_n) - \cdots - \lambda_{n-1}(x_{n-1} - x_n)} \mu^n e^{-\mu(x_1 + \cdots + x_n)} \mathrm{d}x_1 \cdots \mathrm{d}x_n \\ &\quad = n \int \mathbb{1}\{0 \leq u_1, \ldots, u_n\} e^{-\lambda_n u_n - \lambda_1 u_1 - \cdots - \lambda_{n-1} u_{n-1}} \mu^n e^{-\mu(u_1 + \cdots + u_{n-1} + n u_n)} \mathrm{d}u_1 \cdots \mathrm{d}u_n \\ &\quad = \mathbb{E}(e^{-\lambda_n E_n/n}) \mathbb{E}(e^{-\lambda_1 E_1}) \cdots \mathbb{E}(e^{-\lambda_{n-1} E_{n-1}}).\end{aligned}$$

6.

7. Les arguments ci-dessous montrent que $\min_{k=1,\ldots,n} E_k$ est égale en distribution à E_1/n, c'est donc une variable exponentielle de paramètre $n\mu$.

A.3 Exercices du Chapitre 3

↪ Exercice 3.1

1. Pour toute réalisation ω, la suite $(M_n(\omega))$ est croissante et converge donc vers $M_\infty(\omega) = \max_{k \geq 1} U_k(\omega)$: on a donc la convergence presque sûre $M_n \overset{\text{p.s.}}{\to} M_\infty$. Par contre, il n'est pas vrai que $M_\infty(\omega) = 1$ pour tout ω : il existe par exemple ω tel que $U_k(\omega) = 1/2$ pour tout k. En revanche, on va maintenant montrer que l'évènement $E = \{\omega : M_\infty(\omega) = 1\}$ a probabilité 1, ce qui prouvera le résultat $\mathbb{P}(M_n \to 1) = 1$ puisque $\mathbb{P}(E) \leq \mathbb{P}(M_n \to 1)$. Par propriété d'une mesure de probabilités, on a (Proposition 1.1.1)

$$\mathbb{P}(M_\infty \leq x) = \mathbb{P}(U_k \leq x \ \forall k \geq 1) = \lim_{n \to \infty} \mathbb{P}(U_k \leq x \ \forall k = 1, \ldots, n).$$

Puisque les (U_k) sont i.i.d. et uniformément réparties sur $[0, 1]$, on a

$$\mathbb{P}(U_k \leq x \ \forall k = 1, \ldots, n) = \mathbb{P}(U_1 \leq x)^n = x^n.$$

On obtient donc $\mathbb{P}(M_\infty \leq x) = 0$ pour tout $x < 1$, ce qui donne en faisant $x \uparrow 1$ que $\mathbb{P}(M_\infty < 1) = 0$ (Proposition 1.1.1 ou Proposition 1.3.1) et donc $\mathbb{P}(M_\infty = 1) = 1$.

↪ Exercice 3.2

1. On utilise la définition de convergence en loi vue en amphi : par hypothèse on a que $F_{X_n}(x) \to F_c(x)$ pour tout x où F_X est continue. Pour une variable aléatoire constante, on a $F_c(x) = \mathbb{P}(c \leq x) = 1$ si $x \geq c$ et 0 sinon, donc F_X est continue partout sauf en c. Pour montrer que $X_n \overset{\mathbb{P}}{\to} c$ il faut montrer que $\mathbb{P}(|X_n - c| \geq \varepsilon) \to 0$ pour tout $\varepsilon > 0$. On a

$$\mathbb{P}(|X_n - c| \geq \varepsilon) = \mathbb{P}(X_n \geq c + \varepsilon \ \text{ ou } \ X_n \leq c - \varepsilon) = 1 - F_{X_n}(c + \varepsilon) - F_{X_n}(c - \varepsilon).$$

Puisque $c \pm \varepsilon \neq c$ pour $\varepsilon > 0$, on a $F_{X_n}(c + \varepsilon) \to 1$ et $F_{X_n}(c - \varepsilon) \to 0$ ce qui donne le résultat.

2. La Proposition 3.4.4 implique la convergence jointe $(X_n, X_n - Y_n) \overset{\text{L}}{\to} (X, 0)$ et puisque la fonction $f : (x, d) \mapsto x - d$ est continue, on obtient en appliquant la Proposition 3.4.3 que

$$Y_n = f(X_n, X_n - Y_n) \overset{\text{L}}{\to} f(X, 0) = X.$$

Exercice 3.3 *(Convergence jointe)*

1. Si $X+Y \geq \varepsilon$, alors $X \geq \varepsilon/2$ ou $Y \geq \varepsilon/2$, et puisque $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$ on obtient

$$\mathbb{P}(X+Y \geq \varepsilon) \leq \mathbb{P}(X \geq \varepsilon/2 \text{ ou } Y \geq \varepsilon/2) \leq \mathbb{P}(X \geq \varepsilon/2) + \mathbb{P}(Y \geq \varepsilon/2).$$

2. On a

$$\mathbb{P}(|X_n + Y_n| \geq \varepsilon) \leq \mathbb{P}(|X_n| \geq \varepsilon/2) + \mathbb{P}(|Y_n| \geq \varepsilon/2)$$

et chaque terme tend bien vers 0 par définition de la convergence en probabilités. Pour montrer que $(X_n, Y_n) \xrightarrow{\mathbb{P}} (X, Y)$ il faut montrer que

$$\mathbb{P}(\|(X_n, Y_n) - (X, Y)\| \geq \varepsilon) \to 0$$

pour n'importe quelle norme $\|\cdot\|$ sur $\mathbb{R}^2$ (cf. Section 3.5). Si on prend la norme L_1, on a

$$\mathbb{P}(\|(X_n, Y_n) - (X, Y)\| \geq \varepsilon) = \mathbb{P}(|X_n - X| + |Y_n - Y| \geq \varepsilon)$$

et puisque $|X_n - X| \xrightarrow{\mathrm{L}} 0$ et $|Y_n - Y| \xrightarrow{\mathrm{L}} 0$ on obtient bien le résultat.

3. Soit X et X' i.i.d. : alors $X \xrightarrow{\mathrm{L}} X$ et $X \xrightarrow{\mathrm{L}} X'$ mais (X, X) ne converge pas en loi vers (X, X').

↪ **Exercice 3.4** *(Valeurs extrêmes)*

1. $\mathbb{P}(M_n \leq x) = \mathbb{P}(X_1 \leq x, \ldots, X_n \leq x)$: puisque les X_k sont indépendantes, on a $\mathbb{P}(X_1 \leq x, \ldots, X_n \leq x) = \mathbb{P}(X_1 \leq x) \cdots \mathbb{P}(X_n \leq x)$ et puisqu'elles ont même loi, toutes les probabilités apparaissant dans le produit sont égales à $\mathbb{P}(X \leq x)$ ce qui donne le résultat.

2. *(Exponentiel = cas spécial de Weibull)* Par l'exercice 1, on a $M_n \xrightarrow{\text{p.s.}} b$ et puisque convergence presque sûre implique convergence en loi on a aussi $M_n \xrightarrow{\mathrm{L}} b$. Par ailleurs, pour $x \geq 0$ fixé, on a pour n suffisamment grand

$$\mathbb{P}(n(b - M_n) \geq x) = \mathbb{P}(M_n \leq b - x/n) = \mathbb{P}(X \leq b - x/n)^n = \left(1 - \frac{x}{n(b-a)}\right)^n \underset{n\to\infty}{\longrightarrow} e^{-x/(b-a)}.$$

On reconnaît la fonction de répartition d'une loi exponentielle de paramètre $1/(b-a)$.

3. *(Weibull)* On calcule alors pour $x \geq 0$

$$\begin{aligned}
\mathbb{P}(n^{1/2}(1 - M_n) \geq x) &= \mathbb{P}(M_n \leq 1 - x/n^{1/2}) \\
&= \mathbb{P}(X \leq 1 - x/n^{1/2})^n \\
&= \left(1 - \int_{1-x/n^{1/2}}^{1} f(y)\mathrm{d}y\right)^n \\
&= \exp\left(n \log\left(1 - \int_{1-x/n^{1/2}}^{1} f(y)\mathrm{d}y\right)\right) \\
&\underset{n\to\infty}{\longrightarrow} \exp\left(-\frac{\alpha(\alpha+1)}{2}x^2\right)
\end{aligned}$$

puisque les équivalences $\int_{1-\varepsilon}^{1} f(x)\mathrm{d}x \sim \frac{\alpha(\alpha+1)}{2}\varepsilon^2$ et $\log(1+\varepsilon) \sim \varepsilon$ impliquent que

$$n \log\left(1 - \int_{1-x/n^{1/2}}^{1} f(y)\mathrm{d}y\right) \sim -n \int_{1-x/n^{1/2}}^{1} f(y)\mathrm{d}y \sim -n\frac{\alpha(\alpha+1)}{2}\frac{x^2}{n}.$$

Si M_∞ est la variable aléatoire limite, on a donc prouvé que

$$\mathbb{P}(M_\infty \geq x) = \exp\left(-\frac{\alpha(\alpha+1)}{2}x^2\right)$$

et donc en dérivant, on obtient que M_∞ est absolument continue de densité $\alpha(\alpha+1)xe^{-\alpha(\alpha+1)x^2/2}\mathbb{1}\{x \geq 0\}$.

4. *(Gumbel)* Si $u_n = (1/\lambda)\log n$, on a

$$\mathbb{P}(M_n - u_n \leq x) = F_X(x+u_n)^n = \left(1 - e^{-\lambda(x+u_n)}\right)^n = \left(1 - \frac{1}{n}e^{-\lambda x}\right)^n \underset{n\to\infty}{\longrightarrow} \exp\left(-e^{-\lambda x}\right).$$

5. *(Fréchet)* On a

$$\mathbb{P}(M_n/n^{1/\alpha} \leq x) = \left(1 - \frac{1}{(1+xn^{1/\alpha})^\alpha}\right)^n \underset{n\to\infty}{\longrightarrow} \exp\left(-x^{-\alpha}\right).$$

↪ **Exercice 3.5** *(Physique statistique)*

1.

2. Cela suit directement du fait que pour tout $x \in \mathbb{R}$, on a $\lfloor x/\varepsilon \rfloor \sim x/\varepsilon$ lorsque $\varepsilon \to 0$.

3. On a

$$\mathbb{E}(e^{itX_\varepsilon}) = \frac{\varepsilon \sum_{k\in\mathbb{Z}} g(\varepsilon k)}{\varepsilon \sum_{k\in\mathbb{Z}} f(\varepsilon k)}$$

avec $g(x) = e^{itx}f(x)$ et par somme de Riemann, $\varepsilon\sum_{k\in\mathbb{Z}} g(\varepsilon k) \to \int_{\mathbb{R}} g = \mathbb{E}(e^{itX})$ et $\varepsilon\sum_{k\in\mathbb{Z}} f(\varepsilon k) \to \int_{\mathbb{R}} f = 1$, ce qui prouve la convergence des fonctions caractéristiques.

Exercice 3.6 *(Lien avec l'analyse fonctionnelle)*

1. Les deux premières propriétés de l'unité approchée pour la convolution définissent une densité de probabilité, et si X_n est une variable aléatoire de densité U_n, alors la troisième propriété veut exactement dire que $X_n \overset{\mathbb{P}}{\to} 0$. La première question est évidente lorsque l'on reconnaît la densité gaussienne. En terme probabiliste, on a

$$f * U_n(x) = \int f(x-t)U_n(t)\mathrm{d}t = \mathbb{E}(f(x - X_n))$$

et donc

$$\|f * U_n\|_1 \leq \int \mathbb{E}(|f(x-X_n)|)\mathrm{d}x = \mathbb{E}(\int |f(x-X_n)\mathrm{d}x|) = \|f\|_1.$$

Par ailleurs, puisque $X_n \overset{\mathbb{P}}{\to} 0$, on a $f(x-X_n) \overset{\mathbb{P}}{\to} f(x)$ par continuité, puis $\mathbb{E}(f(x-X_n)) \to f(x)$ par domination.

Problème 3.7

1. Il faut $p_k \to 0$ puisque $\mathbb{P}(X_k \neq 0) = p_k$.

2. La suite N_n est presque sûrement croissante.

3. Puisque X_n est à valeurs dans $\{0,1\}$, on a $X_n \overset{\text{p.s.}}{\to} 0$ si et seulement si, presque sûrement, $X_n = 0$ pour n suffisamment grand, ce qui équivaut à $N_\infty < \infty$.

4. On a

$$\mathbb{E}(N_\infty) = \sum_{k\geq 1} \mathbb{E}(X_k) = \sum_{k\geq 1} p_k.$$

Si $\sum_k p_k < \infty$, alors N_∞ est de moyenne finie et est donc presque sûrement finie. On a donc refait la preuve du lemme de Borel–Cantelli.

5. Puisque $N_n \overset{\text{p.s.}}{\to} N_\infty$, on a $N_n \overset{\text{L}}{\to} N_\infty$ et donc $\mathbb{P}(N_n \leq K) \to \mathbb{P}(N_\infty \leq K)$. Ainsi, si $\mathbb{P}(N_n \leq K) \to 0$ pour tout K, cela implique que $\mathbb{P}(N_\infty \leq K) = 0$ pour tout K, et donc $\mathbb{P}(N_\infty = \infty) = 1$ et donc X_n ne converge pas presque sûrement vers 0.

6. On a

$$\mathbb{P}(N_n \leq K) = \mathbb{P}(e^{-N_n} \geq e^{-K}) \leq e^K\mathbb{E}(e^{-N_n}) = e^K\mathbb{E}(e^{-(X_1+\cdots+X_n)}) = e^K\prod_{k=1}^n \mathbb{E}(e^{-X_k})$$

ce qui donne le résultat puisque $\mathbb{E}(e^{-X_k}) = e^{-1}p_k + 1 - p_k$. On voit donc que si $\sum_k p_k = \infty$, alors $\mathbb{P}(N_n \leq K) \to 0$ et donc X_n ne converge pas presque sûrement vers 0 par la question précédente.

7. Pour $p_k = 1/k$, on a donc $p_k \to 0$ mais $\sum_k p_k = \infty$: la suite (X_n) converge en loi vers 0 mais pas presque sûrement !

8. On a $X'_n \overset{\text{p.s.}}{\to} 0$, et ce alors que X'_n est une variable aléatoire de Bernoulli de paramètre $1/n$. La différence avec la suite X_n est que les X'_n ne sont pas indépendantes !

9. Idem.

10. On a

$$X_n \overset{\text{p.s.}}{\to} 0 \iff \mathbb{P}\left(\bigcap_{k\in\mathbb{N}^*} E_k\right) = 1$$

avec E_k l'évènement $E_k = \{N_\infty(1/k) < \infty\}$. Puisque la suite E_k est décroissante, on a

$$\mathbb{P}\left(\bigcap_{k\in\mathbb{N}^*} E_k\right) = \lim_{k\to\infty} \mathbb{P}(E_k)$$

et puisque la suite $\mathbb{P}(E_k)$ à valeurs dans $[0,1]$ est décroissante, on a $P(E_k) \to 1$ si et seulement si $\mathbb{P}(E_k) = 1$ pour chaque k.

11. On utilise le cas Bernoulli pour les variables $\mathbb{1}\{X_k \geq \varepsilon\}$ qui sont Bernoulli de paramètre $e^{-\lambda_k \varepsilon}$.

12. $\lambda_k = \log k$: alors $\lambda_k \to \infty$ (CNS pour que $X_k \overset{\text{L}}{\to} 0$) mais $\sum_k e^{-\lambda_k} = \sum_k (1/k) = \infty$.

A.4 Exercices du Chapitre 4

↪ Exercice 4.1

1. On définit $\mathbf{X}' = \begin{pmatrix} X_1 & X_2 \end{pmatrix}^T$: alors $\mathbf{X}'$ est centrée, et en outre

$$\mathbb{V}\mathrm{ar}(\mathbf{X}') = \begin{pmatrix} 3 & 6 \\ 6 & 14 \end{pmatrix}$$

Le déterminant vaut 6, $\mathbb{V}\mathrm{ar}(\mathbf{X}')$ est donc inversible et donc le théorème 4.5.1 donne

$$\mathbb{E}(X_3 \mid X_1, X_2) = \mathbb{C}\mathrm{ov}(X_3, \mathbf{X}')\mathbb{V}\mathrm{ar}(\mathbf{X}')^{-1}\mathbf{X}'.$$

On calcule

$$\mathbb{V}\mathrm{ar}(\mathbf{X}')^{-1} = \frac{1}{6}\begin{pmatrix} 14 & -6 \\ -6 & 3 \end{pmatrix}$$

et par définition,

$$\mathbb{C}\mathrm{ov}(X_3, \mathbf{X}') = \begin{pmatrix} \mathbb{V}\mathrm{ar}(\mathbf{X})_{31} & \mathbb{V}\mathrm{ar}(\mathbf{X})_{32} \end{pmatrix} = \begin{pmatrix} 2 & 3 \end{pmatrix}$$

ce qui donne

$$\mathbb{E}(X_3 \mid X_1, X_2) = \begin{pmatrix} 2 & 3 \end{pmatrix} \frac{1}{6}\begin{pmatrix} 14 & -6 \\ -6 & 3 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \frac{1}{6}\begin{pmatrix} 2 & 3 \end{pmatrix}\begin{pmatrix} 14X_1 - 6X_2 \\ -6X_1 + 3X_2 \end{pmatrix} = \frac{1}{6}(10X_1 - 3X_2).$$

2. On peut invoquer directement la Proposition 1.7.4, ou alors calculer directement

$$\mathbb{E}(X_i X_n) = \mathbb{E}(\mathbb{E}(X_i X_n \mid X_1, \ldots, X_{n-1})) = \mathbb{E}(X_i \mathbb{E}(X_n \mid X_1, \ldots, X_{n-1}))$$

en utilisant le théorème de l'espérance totale pour la première égalité et la Proposition 1.7.2 pour la seconde.

3. On cherche $\mathbb{E}(X_n \mid X_1, \ldots, X_{n-1}) = \sum_{k=1}^{n-1} \alpha_k X_k$: si cette relation est satisfaite, alors on obtient en utilisant la question précédente pour chaque $i = 1, \ldots, n-1$

$$\mathbb{E}(X_i X_n) = \mathbb{E}\left(X_i \sum_{k=1}^{n-1} \alpha_k X_k\right) = \sum_{k=1}^{n-1} \alpha_k \mathbb{E}(X_i X_k)$$

ce qui donne bien $n-1$ équations satisfaites par les α_k.

4. Pour $n=3$, cela donne

$$\mathbb{E}(X_1X_3) = 2 = \alpha_1\mathbb{E}(X_1^2) + \alpha_2\mathbb{E}(X_1X_2) = 3\alpha_1 + 6\alpha_2$$

et

$$\mathbb{E}(X_2X_3) = 3 = \alpha_1\mathbb{E}(X_2X_1) + \alpha_2\mathbb{E}(X_2^2) = 6\alpha_1 + 14\alpha_2$$

ce qui donne $2\alpha_2 = -1$ et $3\alpha_1 = 5$

↪ **Exercice 4.2** *(Coordonnées polaires)*

1. On utilise le théorème de changement de variables pour des variables aléatoires absolument continues (théorème 2.5.1). On a $(R,\Theta) = \varphi(X,Y)$ avec $\varphi^{-1}(r,\theta) = (r\cos\theta, r\sin\theta)$ pour $(r,\theta) \in (0,\infty)\times(0,2\pi)$, et donc

$$\mathrm{Jac}_{(r,\theta)}(\varphi^{-1}) = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}$$

et donc

$$\det(\mathrm{Jac}_{(r,\theta)}(\varphi^{-1})) = r\cos\theta^2 + r\sin\theta^2 = r$$

ce qui donne

$$\begin{aligned} f_{R,\Theta}(r,\theta) &= r f_{X,Y}(r\cos\theta, r\sin\theta)\mathbb{1}\{(r,\theta)\in[0,\infty)\times[0,2\pi]\} \\ &= r f_X(r\cos\theta) f_Y(r\sin\theta)\mathbb{1}\{(r,\theta)\in[0,\infty)\times[0,2\pi]\} \end{aligned}$$

et donc, pour $r \geq 0$,

$$f_R(r) = \int_0^{2\pi} f_{R,\Theta}(r,\theta)\mathrm{d}\theta = r\int_0^{2\pi} f_X(r\cos\theta) f_Y(r\sin\theta)\mathrm{d}\theta.$$

2. Dans le cas gaussien, cela donne (pour $r\geq 0$ et $\theta\in[0,2\pi]$, sinon la densité vaut 0)

$$f_{R,\Theta}(r,\theta) = r\frac{1}{2\pi}\exp\left(-\frac{1}{2}(r\cos\theta)^2 - \frac{1}{2}(r\sin\theta)^2\right) = r\frac{1}{2\pi}\exp\left(-\frac{1}{2}r^2\right)$$

et donc $f_R(r) = re^{-r^2/2}\mathbb{1}\{r\geq 0\}$, Θ suit une loi uniforme sur $[0,2\pi]$ et R et Θ sont indépendantes.

↪ **Exercice 4.3** *(Loi du χ^2)*

1. Par définition,

$$L_r(t) = \mathbb{E}\left(e^{-t(Y_1^2+\cdots+Y_r^2)}\right)$$

où les Y_i sont i.i.d. standard normaux. Puisque les Y_i sont i.i.d.,

$$L_r(t) = \left[\mathbb{E}\left(e^{-tY_1^2}\right)\right]^r = L_1(t)^r.$$

Le théorème de transfert donne

$$L_1(t) = \frac{1}{\sqrt{2\pi}}\int e^{-tx^2}e^{-x^2/2}\mathrm{d}x = \frac{1}{\sqrt{1+2t}}$$

par changement de variable pour la deuxième égalité, ce qui donne finalement

$$L_r(t) = \frac{1}{(1+2t)^{r/2}}.$$

2. La densité de X est donnée par (Théorème 4.3.4)

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n\det(\mathbb{V}\mathrm{ar}(X))}}\exp\left(-\frac{1}{2}x^T\mathbb{V}\mathrm{ar}(X)^{-1}x\right),$$

et donc le théorème de transfert donne

$$\begin{aligned} L_{X\mathbb{V}\mathrm{ar}(X)^{-1}X^T}(t) &= \mathbb{E}\left(\exp\left(-tX\mathbb{V}\mathrm{ar}(X)^{-1}X^T\right)\right) \\ &= \frac{1}{\sqrt{(2\pi)^n \det(\mathbb{V}\mathrm{ar}(X))}} \int \exp\left(-tx\mathbb{V}\mathrm{ar}(X)^{-1}x^T - \frac{1}{2}x^T\mathbb{V}\mathrm{ar}(X)^{-1}x\right) \mathrm{d}x \\ &= \frac{1}{\sqrt{(2\pi)^n \det(\mathbb{V}\mathrm{ar}(X))}} \int \exp\left(-\frac{1}{2}(1+2t)x\mathbb{V}\mathrm{ar}(X)^{-1}x^T\right) \mathrm{d}x \\ &= \frac{1}{(1+2t)^{n/2}} \int \frac{1}{\sqrt{(2\pi)^n \det(\mathbb{V}\mathrm{ar}(X))}} \exp\left(-\frac{1}{2}x\mathbb{V}\mathrm{ar}(X)^{-1}x^T\right) \mathrm{d}x \end{aligned}$$

où la dernière égalité vient du changement de variable $y = (1+2t)^{1/2}x$. La dernière intégrale vaut 1 puisque c'est l'intégrale de la densité gaussienne standard, ce qui donne bien le résultat.

3. On reconnait la transformée de Laplace d'une loi du χ^2, et puisque la transformée de Laplace caractérise la loi cela montre le résultat.

Exercice 4.4

1. Le théorème de transfert donne $\mathbb{E}(X_i) = \int_0^{2\pi} \cos\theta \frac{1}{2\pi} \mathrm{d}\theta$ qui vaut 0 par symétrie, et le même raisonnement donne $\mathbb{E}(Y_i) = \mathbb{E}(X_iY_i) = 0$. Puisque $\mathbb{C}\mathrm{ov}(X_i, Y_i) = \mathbb{E}(X_iY_i) - \mathbb{E}(X_i)\mathbb{E}(Y_i)$ cela donne $\mathbb{C}\mathrm{ov}(X_i, Y_i) = 0$.

2. La relation $X_i^2 + Y_i^2 = 1$ empêche X_i et Y_i d'être indépendantes. Plus formellement, si X_i et Y_i étaient indépendantes, alors (X_i, Y_i) serait absolument continu (Théorème 2.6.5) et $X_i^2 + Y_i^2$ le serait aussi (Théorème 2.5.1) et on aurait alors $\mathbb{P}(X_i^2 + Y_i^2 = 1) = 0$.

3. Puisque $\mathbb{E}(X_i) = 0$, on a $\mathbb{E}(\bar{X}_n) = \frac{1}{n}(\mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n)) = 0$ et de même $\mathbb{E}(Y_i) = 0$, et donc

$$\mathbb{C}\mathrm{ov}(\bar{X}_n, \bar{Y}_n) = \mathbb{E}(\bar{X}_n\bar{Y}_n) = \frac{1}{n}\mathbb{E}\left(\sum_{i,j=1}^n X_iY_j\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(X_iY_i) + \frac{1}{n}\sum_{i\neq j} \mathbb{E}(X_i)\mathbb{E}(Y_j) = 0$$

4. Le Théorème central limite multi-dimensionnel (Théorème 4.2.1) donne

$$(\bar{X}_n, \bar{Y}_n) \xrightarrow{\mathrm{L}} (X_\infty, Y_\infty)$$

où (X_∞, Y_∞) est un vecteur gaussien centré de matrice de covariance $\mathbb{V}\mathrm{ar}(X_1, Y_1)$. Puisque (X_∞, Y_∞) est un vecteur et que sa matrice de covariance est diagonale (par la question 1), il s'ensuit que X_∞ et Y_∞ sont bien indépendantes (Proposition 4.1.2) !

Exercice 4.5

1. On a

$$\mathbb{C}\mathrm{ov}(X_1 + X_2, X_1 - X_2) = \mathbb{V}\mathrm{ar}(X_1) - \mathbb{V}\mathrm{ar}(X_2) = 0.$$

Puisque $(X_1 + x_2, X_1 - X_2)$ est un vecteur gaussien, décorrélation implique indépendance.

2. On rappelle que $xMy = \sum_{i,j} M_{ij}x_iy_j$. On a

$$\bar{X}_n^2 = \frac{1}{n^2}\sum_{i,j} X_iX_j = \frac{1}{n}X^TA_1X$$

et

$$(n-1)S_{n-1}^2 = \sum_{k=1}^n (X_k - \bar{X}_n)^2 = \sum_{k=1}^n X_k^2 - 2\bar{X}_n\sum_k X_k + n\bar{X}_n^2 = X^TX - n\bar{X}_n^2 = X^TA_2X$$

3. A_1 et A_2 sont symétriques, A_1 est de rang 1 et A_2 de rang $n-1$, et la seule valeur propre non nulle de A_1 est 1.

4. Puisque toute transformation affine d'un vecteur gaussien est un vecteur gaussien, UX est bien un vecteur gaussien. Par ailleurs, il est centré puisque $\mathbb{E}(UX) = U\mathbb{E}(X)$ et que X l'est, et il est bien standard car

$$\mathbb{V}\mathrm{ar}(UX) = U\mathbb{V}\mathrm{ar}(X)U^T = UIU^T = UU^T = I.$$

Utilisant la première question, on voit que

$$\bar{X}_n^2 = nX^T A_1 X = n(UX)^T \Delta (UX) = n(UX)_1^2$$

avec $(UX)_k$ la k-ième coordonnée du vecteur UX, et

$$(n-1)S_{n-1}^2 = X^T A_2 X = (UX)^T (I - \Delta)(UX) = \sum_{k=2}^{n} (UX)_k^2.$$

Cela prouve tout ce qu'il faut :

- $\bar{X}_n$ et S_{n-1}^2 sont bien indépendantes car $\bar{X}_n$ est une fonction de $(UX)_1$ et S_{n-1}^2 de $((UX)_k, k \geq 2)$ et que les $(UX)_k$ sont indépendantes ;
- $\bar{X}_n$ et S_{n-2}^2 suivent bien les lois annoncées.

A.5 Exercices du Chapitre 5

↪ **Exercice 5.1** *(Estimateurs de la moyenne)*

1. Il faut $\mathbb{E}(\sum_{k=1}^n a_k X_k) = \mathbb{E}(X_1)$ soit $\sum_{k=1}^n a_k = 1$.

2. On a

$$\mathbb{V}\mathrm{ar}\left(\bar{X}_n\right) = \frac{1}{n}\mathbb{V}\mathrm{ar}(X_1)$$

et

$$\mathbb{V}\mathrm{ar}\left(\sum_{k=1}^{n} a_k X_k\right) = \mathbb{V}\mathrm{ar}(X_1) \sum_{k=1}^{n} a_k^2.$$

3. On cherche à résoudre le problème

$$\min \sum_{k=1}^{n} a_k^2 \quad \text{sous la contrainte} \quad \sum_{k=1}^{n} a_k = 1.$$

On considère le lagrangien

$$L(a_1, \ldots, a_n, \lambda) = \sum_{k=1}^{n} a_k^2 + \lambda \left(1 - \sum_{k=1}^{n} a_k\right).$$

On a $\partial_i L = 2a_i - \lambda$ et $\partial_\lambda L = 1 - \sum_{k=1}^n a_k$ ce qui donne bien $a_i = 1/n$. On a

$$\mathbb{V}\mathrm{ar}\left(\bar{X}_n\right) = \frac{1}{n}\mathbb{V}\mathrm{ar}(X_1)$$

et

$$\mathbb{V}\mathrm{ar}\left(\sum_{k=1}^{n} a_k X_k\right) = \mathbb{V}\mathrm{ar}(X_1) \sum_{k=1}^{n} a_k^2.$$

↪ **Exercice 5.2** *(Estimateur du maximum de vraisemblance)*

1. On calcule

$$\mathcal{L}(\theta; \mathbf{x}_n) = \frac{1}{\theta^n} \exp\left(\left(\frac{1}{\theta} - 1\right) \sum_{k=1}^{n} \log(1 - x_k)\right)$$

et donc

$$\begin{aligned} \partial_\theta \mathcal{L}(\theta; \mathbf{x}_n) = 0 &\iff \partial_\theta \log \mathcal{L}(\theta; \mathbf{x}_n) = 0 \\ &\iff -\frac{n}{\theta} - \frac{1}{\theta^2} \sum_{k=1}^{n} \log(1 - x_k) \\ &\iff \theta = -\frac{1}{n} \sum_{k=1}^{n} \log(1 - x_k) \end{aligned}$$

ce qui donne finalement

$$\hat{\theta}_n = -\frac{1}{n}\sum_{k=1}^{n}\log(1-X_k).$$

2. On calcule

$$\mathcal{L}(\theta;\mathbf{x}_n) = e^{-n\theta}\prod_{k=1}^{n}\frac{\theta^{x_k}}{x_k!} = \frac{1}{\prod x_k!}\exp\left(-n\theta + \log\theta\sum_{k=1}^{n}x_k\right)$$

et donc en passant à nouveau par la log-vraisemblance $\log\mathcal{L}$ on obtient

$$\partial_\theta\mathcal{L}(\theta;\mathbf{x}_n) = 0 \iff -n + \frac{1}{\theta}\sum_{k=1}^{n}x_k = 0 \iff \theta = \frac{1}{n}\sum_{k=1}^{n}x_k.$$

L'estimateur du maximum de vraisemblance est donc la moyenne empirique.

3. On calcule

$$\mathcal{L}(\theta;\mathbf{x}_n) = \frac{1}{(2\pi)^{n/2}\theta^n}\exp\left(-\frac{1}{2\theta^2}\sum_{k=1}^{n}(x_k-m)^2\right) = \frac{1}{(2\pi)^{n/2}}\exp\left(n\log\theta - \frac{1}{2\theta^2}\sum_{k=1}^{n}(x_k-m)^2\right)$$

et donc en passant à nouveau par la log-vraisemblance $\log\mathcal{L}$ on obtient

$$\partial_\theta\mathcal{L}(\theta;\mathbf{x}_n) = 0 \iff \frac{n}{\theta} - \frac{1}{\theta^3}\sum(x_k-m)^2 = 0 \iff \theta = \left(\frac{1}{n}\sum_{k=1}^{n}(x_k-m)^2\right)^{1/2}.$$

L'estimateur du maximum de vraisemblance est donc donnée par

$$\hat{\theta}_n = \left(\frac{1}{n}\sum_{k=1}^{n}(X_k-m)^2\right)^{1/2}.$$

4. Pour $x_1,\dots,x_n \geq 0$ on a $\mathcal{L}(\theta;\mathbf{x}_n) = \theta^{-n}\mathbb{1}\{\theta \geq \max_k x_k\}$. La vraisemblance vaut 0 pour $\theta < \max_k x_k$, saute à $(\max_k x_k)^{-n}$ en $\theta = \max_k x_k$ puis décroît strictement. L'estimateur du maximum de vraisemblance M_n est donc donnée par $M_n = \max_{k=1,\dots,n} X_k$.

↪ **Exercice 5.3** *(Estimateurs de la loi uniforme)*

1. $\mathbb{P}_\theta$ est la loi uniforme sur $[0,\theta]$.

2. On a $\mathbb{E}_\theta(T_1) = \theta/2$, $\mathbb{E}_\theta(T_1^2) = \theta^{-1}\int_0^\theta x^2\mathrm{d}x = \theta^2/3$ et donc $\mathbb{V}\mathrm{ar}_\theta(T_1) = \theta^2/3 - \theta^2/4 = \theta^2/12$. $2T_1$ est donc bien un estimateur sans biais de θ, néanmoins il ne converge pas.

3. $\mathbb{E}_\theta(\bar{T}_n) = \mathbb{E}(T_1) = \theta/2$ par linéarité et $\mathbb{V}\mathrm{ar}_\theta(\bar{T}_n) = \mathbb{V}\mathrm{ar}(T_1)/n$ (en utilisant $\mathbb{V}\mathrm{ar}(aX+bY) = a^2\mathbb{V}\mathrm{ar}(X) + b^2\mathbb{V}\mathrm{ar}(Y)$ pour X et Y indépendants) et donc $\hat{\theta}_n^{(1)} = 2\bar{T}_n$ est bien un estimateur de θ sans biais, qui plus est convergent puisque $\bar{T}_n \overset{\text{p.s.}}{\to} \mathbb{E}(T_1)$ par la loi forte des grands nombres.

4. Cf. question 4 de l'exercice 2.

5. Pour calculer la loi de M_n on utilise la méthode de la fonction de répartition :

$$\mathbb{P}_\theta(M_n \leq x) = \mathbb{P}_\theta(X_1 \leq x)^n$$

et donc en dérivant, on obtient

$$f_{M_n}(x,\theta) = nf_{X_1}(x,\theta)F_{X_1}(x,\theta)^{n-1}\mathbb{1}\{0 \leq x \leq \theta\} = \frac{nx^{n-1}}{\theta^n}\mathbb{1}\{0 \leq x \leq \theta\}.$$

On en déduit

$$\mathbb{E}_\theta(M_n) = \int x f_{M_n}(x,\theta)\mathrm{d}x = \int_0^\theta \frac{nx^n}{\theta^n}\mathrm{d}x = \frac{n\theta}{n+1},$$

$$\mathbb{E}_\theta(M_n^2) = \int x^2 f_{M_n}(x,\theta)\mathrm{d}x = \int_0^\theta \frac{nx^{n+1}}{\theta^n}\mathrm{d}x = \frac{n\theta^2}{n+2}$$

et donc

$$\mathbb{V}\mathrm{ar}_\theta(M_n) = \frac{n\theta^2}{n+2} - \left(\frac{n\theta}{(n+1)}\right)^2 = \frac{n\theta^2}{(n+1)^2(n+2)}.$$

L'estimateur M_n a donc un biais mais il est convergent.

6. On a

$$\frac{\mathbb{V}\mathrm{ar}_\theta(M_n)}{\mathbb{V}\mathrm{ar}_\theta(\bar{T}_n)} = \frac{12n^2}{(n+1)^2(n+2)} \underset{n\to\infty}{\longrightarrow} 0$$

et donc M_n converge a priori beaucoup plus vite que $\bar{T}_n$.

↪ **Exercice 5.4** *(Estimateur de la loi de Bernoulli)*

1. On considère le modèle paramétrique $\{\mathbb{P}_\theta : \theta \in [0,1]\}$ avec $\mathbb{P}_\theta$ la loi binomiale de paramètre θ. On a

$$\mathcal{L}(\theta; \mathbf{x}_n) = \theta^{\sum_k x_k}(1-\theta)^{n-\sum_k x_k}$$

et donc (en considérant encore la log-vraisemblance par exemple)

$$\partial_\theta \mathcal{L}(\theta; \mathbf{x}_n) = 0 \iff \frac{1}{\theta}\sum_k x_k - \frac{1}{1-\theta}\left(n - \sum_k x_k\right) = 0 \iff \theta = \frac{1}{n}\sum_{k=1}^n x_k$$

ce qui donne pour l'estimateur du maximum de vraisemblance

$$\hat{\theta} = \frac{1}{n}\sum_{k=1}^n X_i$$

dont la variance vaut $\mathbb{V}\mathrm{ar}(\hat{\theta}) = \mathbb{V}\mathrm{ar}(X_1)/n$.

2. Puisque $\hat{\theta}^{(i)}$ sont des moyennes empiriques, ils sont sans biais et donc leur moyenne aussi. On cherche a_1 et a_2 tels que $a_1 + a_2 = 1$ (pour que l'estimateur soit sans biais) et de variance minimale. On calcule

$$\mathbb{V}\mathrm{ar}(a_1\hat{\theta}^{(1)} + a_2\hat{\theta}^{(2)}) = a_1^2\mathbb{V}\mathrm{ar}(\hat{\theta}^{(1)}) + a_2^2\mathbb{V}\mathrm{ar}(\hat{\theta}^{(2)}).$$

Le lagrangien vaut donc

$$L(a_1, a_2, \lambda) = a_1^2\mathbb{V}\mathrm{ar}(\hat{\theta}^{(1)}) + a_2^2\mathbb{V}\mathrm{ar}(\hat{\theta}^{(2)}) + \lambda(1 - a_1 - a_2)$$

et donc $\partial_i L = 2a_i\mathbb{V}\mathrm{ar}(\hat{\theta}^{(i)}) - \lambda$ et $\partial_\lambda L = 1 - a_1 - a_2$ ce qui donne comme a_1 et a_2 optimaux

$$a_1 = \frac{\mathbb{V}\mathrm{ar}(\hat{\theta}^{(2)})}{\mathbb{V}\mathrm{ar}(\hat{\theta}^{(1)}) + \mathbb{V}\mathrm{ar}(\hat{\theta}^{(2)})} = \frac{n_1}{n_1+n_2} = 1 - a_2.$$

Dans la classe des estimateurs de la forme $a_1\hat{\theta}^{(1)} + a_2\hat{\theta}^{(2)}$, trouvez par le calcul celui sans biais et de variance minimale.

3. On a

$$a_1\hat{\theta}^{(1)} + a_2\hat{\theta}^{(2)} = \frac{a_1}{n_1}\sum_{k=1}^{n_1} X_k^{(1)} + \frac{a_2}{n_2}\sum_{k=1}^{n_2} X_k^{(2)}$$

et donc d'après l'exercice 1, il faut prendre $a_1 + a_2 = 1$ et $a_1/n_1 = a_2/n_2$.

↪ **Exercice 5.5** *(Loi de Bernoulli)*

1. On a

$$\mathbb{E}_\theta(\hat{\theta}_n) = \mathbb{E}_\theta\left(\frac{1}{n}\sum_{k=1}^n X_k\right) = \mathbb{E}_\theta(X_1) = \theta.$$

$\hat{\theta}_n$ est donc sans biais, sa variance est donc égale à l'erreur quadratique moyenne et

$$\mathbb{V}\mathrm{ar}_\theta(\hat{\theta}_n) = \mathbb{V}\mathrm{ar}_\theta\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = \frac{1}{n}\mathbb{V}\mathrm{ar}_\theta(X_1) = \frac{\theta(1-\theta)}{n}.$$

$\hat{\theta}_n$ est convergent presque sûrement par la loi forte des grands nombres. On vérifie maintenant qu'il est efficace en calculant la borne de Fréchet–Darmois–Cramer–Rao : $p(x;\theta) = \theta^x(1-\theta)^{1-x}$

$$\log p(x;\theta) = x\log\theta + (1-x)\log(1-\theta)$$

et donc

$$\partial_\theta^2 \log p(x;\theta) = -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}$$

et donc

$$I(\theta) = -\mathbb{E}_\theta(\partial_\theta^2 \log p(X_1;\theta)) = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}.$$

$\hat{\theta}_n$ est donc efficace.

2. Par le théorème central limite, $\sqrt{n}(\hat{\theta}_n - \theta)/\sqrt{\mathbb{V}\mathrm{ar}_\theta(X_1)}$ sous $\mathbb{P}_\theta$ converge en loi vers une loi normale standard (cf. Proposition 3.4.6 du poly de proba) : on retrouve donc le théorème 5.6.2 du cours.

3. Appelons φ la fonction décroissante de l'énoncé : puisqu'elle est continue, on a donc que $\varphi^{-1}(I)$ reste un intervalle. On montre maintenant que c'est un intervalle de confiance de niveau asymptotique $F(a_+)-F(a_-)$: on a

$$\mathbb{P}_\theta\left(\theta \in I(\hat{\theta}_n, a_+/\sqrt{n}, a_-/\sqrt{n})\right) = \mathbb{P}_\theta\left(\frac{a_-}{\sqrt{n}} \le \frac{\hat{\theta}_n - \theta}{\sqrt{\theta(1-\theta)}} \le \frac{a_+}{\sqrt{n}}\right) = \mathbb{P}_\theta\left(a_- \le \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sqrt{\mathbb{V}\mathrm{ar}_\theta(\hat{\theta}_n)}} \le a_+\right)$$

qui tend bien par la question précédente vers $F(a_+) - F(a_-)$.

4. A FAIRE

↪ **Exercice 5.6** *(Loi beta)*

1. On a

$$\begin{aligned}
\mathbb{P}_\theta\left(-\log(1-X_i) \le x\right) &= \mathbb{P}_\theta\left(1 - X_i \ge e^{-x}\right) \\
&= \mathbb{P}_\theta\left(X_i \le 1 - e^{-x}\right) \\
&= \int_0^{1-e^{-x}} \frac{1}{\theta}(1-y)^{1/\theta-1}\mathrm{d}y \\
&= \frac{1}{\theta}\int_{e^{-x}}^{1} y^{1/\theta-1}\mathrm{d}y \\
&= 1 - e^{-x/\theta}
\end{aligned}$$

ce qui montre que $-\log(1-X_i)$ suit une loi exponentielle de paramètre $1/\theta$. Sa moyenne et son écart-type valent donc θ.

2. Comme à l'exercice précédent, on calcule $\mathbb{E}_\theta(\hat{\theta}_n) = \mathbb{E}_\theta(-\log(1-X_1)) = \theta$. $\hat{\theta}_n$ est donc sans biais, sa variance et son erreur quadratique moyenne sont donc égales et données par

$$\mathbb{V}\mathrm{ar}_\theta(\hat{\theta}_n) = \frac{1}{n}\mathbb{V}\mathrm{ar}_\theta(-\log(1-X_1)) = \frac{\theta^2}{n}.$$

$\hat{\theta}_n$ est convergent presque sûrement par la loi forte des grands nombres. Pour répondre à la question de l'efficacité, il faut calculer la borne de Fréchet–Darmois–Cramer–Rao : $p(x;\theta) = f_\theta(x)$ et donc

$$\log p(x;\theta) = -\log\theta + \log(1-x)/\theta - \log(1-x)$$

et donc

$$\partial_\theta^2 \log p(x;\theta) = \frac{1}{\theta^2} + \frac{2\log(1-x)}{\theta^3}$$

et donc

$$I(\theta) = -\mathbb{E}_\theta(\partial_\theta^2 \log p(X_1;\theta)) = -\frac{1}{\theta^2} + \frac{2}{\theta^2} = \frac{1}{\theta^2}.$$

$\hat\theta_n$ est donc efficace.

3. On écrit $\hat\theta_n/\theta = \sum_{k=1}^n Y_i$ avec $Y_i = -(n\theta)^{-1}\log(1-X_i)$ pour voir grâce aux indications que $\hat\theta_n/\theta$ suit une loi Gamma(n,n) de fonction de répartition G_n. On utilise alors le fait que cette loi ne dépend pas θ pour construire des intervalles de confiance : ainsi, si $0 < a_- < a_+$ satisfont $G_n(a_+) - G_n(a_-) = 1-\alpha$, alors $[\hat\theta_n/a_+, \hat\theta_n/a_-]$ est un intervalle de confiance au niveau $1-\alpha$ puisque

$$\mathbb{P}_\theta\left(\frac{\hat\theta_n}{a_+} < \theta < \frac{\hat\theta_n}{a_-}\right) = \mathbb{P}_\theta\left(a_- \le \frac{\hat\theta_n}{\theta} \le a_+\right) = G_n(a_+) - G_n(a_-) = 1-\alpha.$$

4. Puisque $\mathbb{V}\mathrm{ar}_\theta(-\log(1-X_1)) = \theta^2$ et $\mathbb{E}_\theta(-\log(1-X_1)) = \theta$, $\sqrt{n}(\hat\theta_n - \theta)/\theta$ sous $\mathbb{P}_\theta$ converge vers une variable standard normale par le théorème central limite (cf. Proposition 3.4.8 du poly). On aurait directement pu obtenir ce résultat grâce théorème 5.6.2 du cours. On en déduit donc que $[\hat\theta_n/(a_+/\sqrt{n}+1), \hat\theta_n/(a_-/\sqrt{n}+1)]$ pour tout $a_- < a_+$ satisfaisant $F(a_+) - F(a_-) = 1-\alpha$ avec F la fonction de répartition de la loi normale standard est un intervalle de confiance asymptotique puisque

$$\mathbb{P}_\theta\left(\frac{\hat\theta_n}{a_+/\sqrt{n}+1} \le \theta \le \frac{\hat\theta_n}{a_-/\sqrt{n}+1}\right) = \mathbb{P}_\theta\left(\frac{a_-}{\sqrt{n}} + 1 \le \frac{\hat\theta_n}{\theta} \le \frac{a_+}{\sqrt{n}} + 1\right) \underset{n\to\infty}{\longrightarrow} F(a_+) - F(a_-).$$

↪ **Exercice 5.7** *(Loi uniforme)*

1. Le support de $\mathbb{P}_\theta$ dépend de θ : le modèle n'est donc pas régulier et on ne peut a priori pas utiliser tous les résultats du cours.

2. On a $\hat\theta_n/\theta = \max_{k=1,\dots,n}(X_k/\theta)$ et donc le résultat découle directement du fait que sous $\mathbb{P}_\theta$, les (X_k/θ) sont i.i.d. et de loi uniforme sur $[0,1]$. On en déduit que $[\hat\theta_n/a_+, \hat\theta_n/a_-]$ est un intervalle de confiance pour θ au niveau $1-\alpha$ pour tout a_-, a_+ satisfaisant $(a_+)^n - (a_-)^n = 1-\alpha$:

$$\mathbb{P}_\theta\left(\frac{\hat\theta_n}{a_+} \le \theta \le \frac{\hat\theta_n}{a_-}\right) = \mathbb{P}_\theta\left(a_- \le \frac{\hat\theta_n}{\theta} \le a_+\right) = \mathbb{P}_1\left(a_- \le \hat\theta_n \le a_+\right) = (a_+)^n - (a_-)^n.$$

3. On a

$$\mathbb{P}_\theta(n(1-\hat\theta_n/\theta) \ge x) = \mathbb{P}_1(\hat\theta_n \le 1 - x/n) = \left(1 - \frac{x}{n}\right)^n \to e^{-x}$$

et donc $n(1-\hat\theta_n/\theta)$ sous $\mathbb{P}_\theta$ converge en loi vers une variable exponentielle de paramètre 1. On montre donc comme précédemment que

$$\left[\frac{\hat\theta_n}{1-a_-/n}, \frac{\hat\theta_n}{1-a_+/n}\right]$$

est un intervalle de confiance de niveau asymptotique $1-\alpha$ pour tous a_-, a_+ satisfaisant $e^{-a_-} - e^{-a_+} = 1-\alpha$.

Problème 5.8 *(Statistique bayésienne)*

1. Si $h(\mathbf{X}_n) = \mathbb{E}(\Theta \mid \mathbf{X}_n)$, on a

$$\begin{aligned}
\mathbb{E}\big[\big(g(\mathbf{X}_n) - \Theta\big)^2 \mid \mathbf{X}_n\big] &= \mathbb{E}\big[\big(g(\mathbf{X}_n) - h(\mathbf{X}_n) + h(\mathbf{X}_n) - \Theta\big)^2 \mid \mathbf{X}_n\big] \\
&= \big(g(\mathbf{X}_n) - h(\mathbf{X}_n)\big)^2 + \mathbb{E}\big[\big(h(\mathbf{X}_n) - \Theta\big)^2 \mid \mathbf{X}_n\big] \\
&\qquad + 2\big(g(\mathbf{X}_n) - h(\mathbf{X}_n)\big)\mathbb{E}\big[h(\mathbf{X}_n) - \Theta \mid \mathbf{X}_n\big] \\
&= \big(g(\mathbf{X}_n) - h(\mathbf{X}_n)\big)^2 + \mathbb{E}\big[\big(h(\mathbf{X}_n) - \Theta\big)^2 \mid \mathbf{X}_n\big].
\end{aligned}$$

Puisque $\big(g(\mathbf{X}_n)-h(\mathbf{X}_n)\big)^2 \geq 0$ on obtient que $\mathbb{E}\big[\big(g(\mathbf{X}_n)-\Theta\big)^2 \mid \mathbf{X}_n\big] \geq \mathbb{E}\big[\big(h(\mathbf{X}_n)-\Theta\big)^2 \mid \mathbf{X}_n\big]$ ce qui donne le résultat en prenant l'espérance et en utilisant le théorème de l'espérance totale.
Première partie. Dans la première partie du problème, on suppose que $\Theta = \theta_0 \in \mathbb{R}$ est une variable déterministe (i.e., une constante).

2. Si $\Theta = \theta_0$ alors $\mathbb{E}(\Theta \mid \mathbf{X}_n) = \theta_0$: l'"estimateur" de Θ est Θ lui-même !

3. On est en train d'estimer la moyenne d'une normale, on a déjà vu que l'estimateur du maximum de vraisemblance est alors la moyenne empirique $\bar{X}_n$ qui converge donc presque sûrement vers $\theta_0 = \Theta$.

4. On calcule

$$\mathrm{EQM}(\bar{\mathrm{X}}_\mathrm{n}) = \mathbb{E}\big[\big(\bar{\mathrm{X}}_\mathrm{n}-\theta_0\big)^2\big] = \mathbb{E}\left[\left(\frac{1}{\mathrm{n}}\sum_{\mathrm{k}=1}^{\mathrm{n}} \mathrm{W}_\mathrm{k}\right)^2\right] = \frac{\sigma_\mathrm{w}^2}{\mathrm{n}}.$$

Cette erreur quadratique moyenne est plus élevée que celle de $\mathbb{E}(\Theta \mid \mathbf{X}_n)$ puisqu'on a prouvé que cette variable aléatoire était d'erreur quadratique moyenne minimale.

5. On a $\bar{X}_n = \Theta + (1/n)\sum_{k=1}^n W_k$. Puisque les W_k sont i.i.d. centrés, la loi forte des grands nombres donne $\bar{X}_n \overset{\text{p.s.}}{\to} \Theta$.

6. $(\Theta, \mathbf{X}_n)$ est obtenu de $(\Theta, W_1, \ldots, W_n)$ qui est un vecteur gaussien (puisque les coordonnées sont indépendantes) par transformation affine, et reste donc un vecteur gaussien. Pour calculer sa matrice de covariance, on calcule

$$\mathbb{C}\mathrm{ov}(\Theta, X_i) = \mathbb{C}\mathrm{ov}(\Theta, \Theta + W_i) = \mathbb{V}\mathrm{ar}(\Theta),\ \mathbb{V}\mathrm{ar}(X_i) = \mathbb{V}\mathrm{ar}(\Theta + W_i) = \mathbb{V}\mathrm{ar}(\Theta) + \sigma_w^2$$

et, pour $i \neq j$,

$$\mathbb{C}\mathrm{ov}(X_i, X_j) = \mathbb{C}\mathrm{ov}(\Theta + W_i, \Theta + W_j) = \mathbb{V}\mathrm{ar}(\Theta)$$

pour obtenir le résultat.

7. Le Théorème 4.5.1 du polycopié donne

$$\mathbb{E}(\Theta \mid \mathbf{X}_n) = \mathbb{E}(\Theta) + \mathbb{C}\mathrm{ov}(\Theta, \mathbf{X}_n)\mathbb{V}\mathrm{ar}(\mathbf{X}_n)^{-1}(\mathbf{X}_n - \mathbb{E}(\mathbf{X}_n)).$$

On vient de voir que $\mathbb{V}\mathrm{ar}(\mathbf{X}_n) = \mathbb{V}\mathrm{ar}(\Theta)H + \sigma_w^2 I$ et donc (d'après l'indication)

$$\mathbb{V}\mathrm{ar}(\mathbf{X}_n)^{-1} = -\frac{\mathbb{V}\mathrm{ar}(\Theta)}{\sigma_w^4 + n\mathbb{V}\mathrm{ar}(\Theta)\sigma_w^2}H + \sigma_w^{-2} I.$$

En outre, $\mathbb{C}\mathrm{ov}(\Theta, \mathbf{X}_n) = \mathbb{V}\mathrm{ar}(\Theta)\mathbf{1}^T$ avec $\mathbf{1}^T = (1 \cdots 1) \in \mathbb{R}^n$. En particulier,

$$\begin{aligned}
\mathbb{C}\mathrm{ov}(\Theta, \mathbf{X}_n)\mathbb{V}\mathrm{ar}(\mathbf{X}_n)^{-1}(\mathbf{X}_n - \mathbb{E}(\mathbf{X}_n)) &= \mathbb{V}\mathrm{ar}(\Theta)\mathbf{1}^T\left(-\frac{\mathbb{V}\mathrm{ar}(\Theta)}{\sigma_w^4 + n\mathbb{V}\mathrm{ar}(\Theta)\sigma_w^2}H + \sigma_w^{-2}I\right)(\mathbf{X}_n - \mathbb{E}(\Theta)\mathbf{1}) \\
&= \mathbb{V}\mathrm{ar}(\Theta)\left(-\frac{\mathbb{V}\mathrm{ar}(\Theta)}{\sigma_w^4 + n\mathbb{V}\mathrm{ar}(\Theta)\sigma_w^2}\mathbf{1}^T H + \sigma_w^{-2}\mathbf{1}^T I\right)(\mathbf{X}_n - \mathbb{E}(\Theta)\mathbf{1}) \\
&= \mathbb{V}\mathrm{ar}(\Theta)\left(-\frac{\mathbb{V}\mathrm{ar}(\Theta)}{\sigma_w^4 + n\mathbb{V}\mathrm{ar}(\Theta)\sigma_w^2}n\mathbf{1}^T + \sigma_w^{-2}\mathbf{1}^T\right)(\mathbf{X}_n - \mathbb{E}(\Theta)\mathbf{1}) \\
&= \frac{\mathbb{V}\mathrm{ar}(\Theta)}{\sigma_w^2}\left(-\frac{n\mathbb{V}\mathrm{ar}(\Theta)}{\sigma_w^2 + n\mathbb{V}\mathrm{ar}(\Theta)} + 1\right)(\mathbf{1}^T\mathbf{X}_n - \mathbb{E}(\Theta)\mathbf{1}^T\mathbf{1}) \\
&= \frac{\mathbb{V}\mathrm{ar}(\Theta)}{\sigma_w^2}\frac{\sigma_w^2}{\sigma_w^2 + n\mathbb{V}\mathrm{ar}(\Theta)}(n\bar{X}_n - n\mathbb{E}(\Theta)) \\
&= \frac{n\mathbb{V}\mathrm{ar}(\Theta)}{\sigma_w^2 + n\mathbb{V}\mathrm{ar}(\Theta)}(\bar{X}_n - \mathbb{E}(\Theta))
\end{aligned}$$

ce qui donne le résultat pour $\mathbb{E}(\Theta \mid \mathbf{X}_n)$.

8. On voit que $\mathbb{E}(\Theta \mid \mathbf{X}_n)$ ne dépend que de la moyenne et de la variance de Θ (en plus des observations X_k) qui doivent donc être connus si l'on veut pouvoir utiliser cet estimateur. Si ces paramètres sont connus, on peut alors interpréter la formule comme une pondération entre les observations et la connaissance a priori $\mathbb{E}(\Theta)$: pour n petit les observations ne participent pas beaucoup et la connaissance a priori domine, pour n grand c'est l'inverse.

9. Soit $\bar{W}_n = (1/n)\sum_{k=1}^n W_k$: on calcule

$$\begin{aligned}\mathrm{EQM}(\mathbb{E}(\Theta \mid \mathbf{X}_\mathrm{n})) &= \mathbb{E}\big[\big(\alpha_n \bar{X}_n + (1-\alpha_n)\mathbb{E}(\Theta) - \Theta\big)^2\big] \\ &= \mathbb{E}\big[\big(\alpha_n \bar{W}_n + \alpha_n\Theta + (1-\alpha_n)\mathbb{E}(\Theta) - \Theta\big)^2\big] \\ &= \alpha_n^2 \mathbb{V}\mathrm{ar}(\bar{W}_n) + (1-\alpha_n)^2\mathbb{V}\mathrm{ar}(\Theta)\end{aligned}$$

ce qui donne le résultat.

10. $\hat{\Theta}_n$ est bien un estimateur de Θ, et on a $\hat{\Theta}_n \overset{\text{p.s.}}{\to} \Theta$ si $a_n \to 1$.

11. On calcule

$$\begin{aligned}\mathrm{EQM}(\hat{\Theta}_\mathrm{n}) &= \mathbb{E}\big[\big(a_n \bar{X}_n + (1-a_n)\mu - \Theta\big)^2\big] \\ &= \mathbb{E}\big[\big(a_n \bar{W}_n + (1-a_n)(\mu - \mathbb{E}(\Theta)) + (1-a_n)(\mathbb{E}(\Theta) - \Theta)\big)^2\big] \\ &= a_n^2 \mathbb{V}\mathrm{ar}(\bar{W}_n) + (1-a_n)^2\mathbb{V}\mathrm{ar}(\Theta) + (1-a_n)^2(\mu - \mathbb{E}(\Theta))^2.\end{aligned}$$

12. Par la question 1, $\mathrm{EQM}(\mathbb{E}(\Theta \mid \mathbf{X}_\mathrm{n})$ est plus petit que $\mathrm{EQM}(\hat{\Theta}_\mathrm{n})$ et $\mathrm{EQM}(\bar{\mathrm{X}}_\mathrm{n})$. Pour comparer ces deux derniers termes, on obtient en utilisant les expressions calculées précédemment que $\mathrm{EQM}(\hat{\Theta}) \leq \mathrm{EQM}(\bar{\mathrm{X}}_\mathrm{n})$ si et seulement si

$$a_n^2 \mathbb{V}\mathrm{ar}(\bar{W}_n) + (1-a_n)^2\mathbb{V}\mathrm{ar}(\Theta) + (1-a_n)^2(\mu - \mathbb{E}(\Theta))^2 \leq \mathbb{V}\mathrm{ar}(\bar{W}_n)$$

ce qui est équivalent à

$$(1-a_n)\mathbb{V}\mathrm{ar}(\Theta) + (1-a_n)(\mu - \mathbb{E}(\Theta))^2 \leq (1+a_n)\frac{\sigma_w^2}{n}.$$

13. Simple calcul à partir de la question 7. L'intérêt est que l'on peut faire une simple mise à jour de $\mathbb{E}(\Theta \mid \mathbf{X}_{n+1})$ à partir de l'ancienne estimation $\mathbb{E}(\Theta \mid \mathbf{X}_n)$ et de la nouvelle observation X_{n+1} sans avoir besoin de tout recalculer.

Problème 5.9

1. Non : les variables T_i ne sont ni indépendantes (elles sont juste deux à deux décorrélées), ni identiquement distribuées.

2. On définit

$$\varphi(\theta) = \varphi(\alpha, \beta) = \sum_{i=1}^n (T_i - \alpha - \beta x_i)^2$$

de sorte que

$$\partial_\alpha \varphi = 2\sum_{i=1}^n (\alpha + \beta x_i - T_i) \quad \text{et} \quad \partial_\beta \varphi = 2\sum_{i=1}^n x_i(\alpha + \beta x_i - T_i).$$

On vérifie bien que $\hat{\beta}_n = C(\mathbf{x}_n, \mathbf{T}_n)/V(\mathbf{x}_n)$ et que $\hat{\alpha}_n = m(\mathbf{T}_n) - \hat{\beta}_n m(\mathbf{x}_n)$ annule le gradient et que la hessienne en ce point est définie positive, il s'agit donc bien de la variable qui minimise l'erreur quadratique moyenne.

3. Par linéarité, $\mathbb{E}(m(\mathbf{T}_n)) = m(\mathbb{E}(\mathbf{T}_n)) = m(\alpha\mathbf{1} + \beta\mathbf{x}_n) = \alpha + \beta m(\mathbf{x}_n)$, avec $\mathbf{1}$ le vecteur avec que des 1, et donc $\mathbb{E}(\hat{\alpha}_n) = \alpha + (\beta - \mathbb{E}(\hat{\beta}_n))m(\mathbf{x}_n)$: il suffit donc de montrer que $\hat{\beta}_n$ est sans biais. Encore par linéarité, $\mathbb{E}(\hat{\beta}_n) = C(\mathbf{x}_n, \mathbb{E}(\mathbf{T}_n))/V(\mathbf{x}_n) = C(\mathbf{x}_n, \alpha\mathbf{1} + \beta\mathbf{x}_n)/V(\mathbf{x}_n)$ et puisque $\alpha\mathbf{1}$ est constant, $C(\mathbf{x}_n, \alpha\mathbf{1} + \beta\mathbf{x}_n) = C(\mathbf{x}_n, \beta\mathbf{x}_n) = \beta C(\mathbf{x}_n, \mathbf{x}_n) = \beta V(\mathbf{x}_n)$ et on obtient bien que $\hat{\theta}_n$ est sans biais.

4. $\mathbf{T}_n$ est un vecteur gaussien de moyenne $\alpha\mathbf{1} + \beta\mathbf{x}_n$ et de matrice de variance-covariance $\sigma^2 I$. On a donc

$$p_n(\mathbf{t}_n; \theta) = \frac{1}{\sigma^n(2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^n (t_k - \alpha - \beta x_k)^2\right).$$

Trouver le maximum de vraisemblance revient à maximiser la log-vraisemblance $\log p_n(\mathbf{x}; \cdot)$, et donc d'après la forme ci-dessus à minimiser l'erreur quadratique.

5. $\hat{\theta}_n$ est obtenu par transformation linéaire des $\mathbf{T}_n$, c'est donc un vecteur gaussien de moyenne (α, β) et dont la matrice de variance-covariance a été admise plus haut. On en déduit donc par exemple que $(\alpha - \hat{\alpha}_n)/\sqrt{\mathbb{V}\mathrm{ar}_\theta(\hat{\alpha}_n)}$ suit une loi normale standard, et donc

$$\left[\hat{\alpha}_n + \sqrt{\mathbb{V}\mathrm{ar}_\theta(\hat{\alpha}_n)}a_-, \hat{\alpha}_n + \sqrt{\mathbb{V}\mathrm{ar}_\theta(\hat{\alpha}_n)}a_+\right] = \left[\hat{\alpha}_n + \frac{a_-\sigma m(\mathbf{x}_n^2)^{1/2}}{\sqrt{(n-1)V(\mathbf{x}_n)}}, \hat{\alpha}_n + \frac{a_+\sigma m(\mathbf{x}_n^2)^{1/2}}{\sqrt{(n-1)V(\mathbf{x}_n)}}\right]$$

un intervalle de confiance pour α de niveau $F(a_+) - F(a_-)$ avec F la fonction de répartition de la loi normale standard.

6. On a maintenant $\theta = (\alpha, \beta, \sigma)$: la vraisemblance est inchangée, mais on calcule sa dérivée par rapport à σ pour trouver

$$\partial_\sigma \log p_n(\mathbf{t}_n; \theta) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{k=1}^{n}(t_k - \alpha - \beta x_k)^2.$$

Puisque la dérivée s'annule pour $\alpha = \hat{\alpha}_n$ et $\beta = \hat{\beta}_n$, on obtient bien le résultat pour $\hat{\sigma}_n$.

7. On reprend les calculs de la question 2 pour obtenir

$$\partial_\alpha \log p_n = \frac{1}{\sigma^2}\sum_{k=1}^{n}(t_k - \alpha - \beta x_k) \quad \text{et} \quad \partial_\beta \log p_n = \frac{1}{\sigma^2}\sum_{k=1}^{n} x_k(t_k - \alpha - \beta x_k)$$

et donc

$$M_n(\theta) = \frac{n}{\sigma^2}\begin{pmatrix} 1 & m(\mathbf{x}_n) \\ m(\mathbf{x}_n) & m(\mathbf{x}_n^2) \end{pmatrix}$$

et donc

$$M_n(\theta)^{-1} = \frac{\sigma^2}{n}\frac{1}{m(\mathbf{x}_n^2) - m(\mathbf{x}_n)^2}\begin{pmatrix} m(\mathbf{x}_n^2) & -m(\mathbf{x}_n) \\ -m(\mathbf{x}_n) & 1 \end{pmatrix} = \frac{\sigma^2}{(n-1)V(\mathbf{x}_n)}\begin{pmatrix} m(\mathbf{x}_n^2) & -m(\mathbf{x}_n) \\ -m(\mathbf{x}_n) & 1 \end{pmatrix}.$$

L'estimateur $\hat{\theta}_n$ est donc efficace.

8. On voit que la largeur de l'intervalle est une fonction de z de la forme

$$z \mapsto c\sqrt{1 + d(z - m(\mathbf{x}_n))^2}$$

avec c et d indépendants de z. L'intervalle est donc le plus petit autour de la moyenne $m(\mathbf{x}_n)$, qui seront donc les valeurs les mieux prédites par le modèle.

A.6 Exercices du Chapitre 6

↪ **Exercice 6.1** *(Risques de première et de deuxième espèce)*

1. Le test est de la forme

$$\text{Accepter } H_0 \iff \bar{X}_n < \kappa.$$

Par ailleurs, on veut choisir n et κ tels que

$$\begin{cases} \alpha = \mathbb{P}_1(\bar{X}_n > \kappa) = 1 - F\left(\frac{(\kappa-1)\sqrt{n}}{\sigma}\right) \\ \eta = \mathbb{P}_2(\bar{X}_n > \kappa) = 1 - F\left(\frac{(\kappa-2)\sqrt{n}}{\sigma}\right) \end{cases} \iff \begin{cases} (\kappa-1)\sqrt{n} = \sigma F^{-1}(1-\alpha) \\ (\kappa-2)\sqrt{n} = \sigma F^{-1}(1-\eta) \end{cases}$$

et donc

$$\sqrt{n} = \sigma F^{-1}(1-\alpha) - \sigma F^{-1}(1-\eta).$$

Pour que le membre de droite soit > 0 il faut $\alpha < \eta$ ce qui est raisonnable. Par ailleurs, une application numérique donne $n \approx 35$.

2. On regarde chaque test.

Cas a$H_1 : \theta = 3$**.** Dans ce cas le test est de la forme $\theta = 2 \Leftrightarrow \bar{X}_n < \kappa$ et on a donc le système

$$\left\{ \begin{array}{l} \alpha = \mathbb{P}(N(2, \sigma/\sqrt{n}) > \kappa) = 1 - F(\sqrt{n}(\kappa - 2)/\sigma) \\ \eta = \mathbb{P}(N(3, \sigma/\sqrt{n}) > \kappa) = 1 - F(\sqrt{n}(\kappa - 3)/\sigma) \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} (\kappa - 2)\sqrt{n} = \sigma F^{-1}(1 - \alpha) \\ (\kappa - 3)\sqrt{n} = \sigma F^{-1}(1 - \alpha) \end{array} \right.$$

On peut résoudre de système, on peut donc répondre aux critèrs imposés.

Cas b : $H_1 : \theta = 1{,}999$**.** Dans ce cas le test est de la forme $\theta = 2 \Leftrightarrow \bar{X}_n > \kappa$ et on a donc le système

$$\left\{ \begin{array}{l} \alpha = \mathbb{P}(N(2, \sigma/\sqrt{n}) < \kappa) = F(\sqrt{n}(\kappa - 2)/\sigma) \\ \eta = \mathbb{P}(N(2 - \varepsilon, \sigma/\sqrt{n}) < \kappa) = F(\sqrt{n}(\kappa - 2 + \varepsilon)/\sigma) \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} \sqrt{n}(\kappa - 2) = \sigma F^{-1}(\alpha) \\ \sqrt{n}(\kappa - 2 + \varepsilon) = \sigma F^{-1}(\eta) \end{array} \right.$$

Encore une fois, on peut répondre aux contraintes, mais on voit que plus ε est petit et plus il faut prendre n grand puisque $\sqrt{n}\varepsilon = \sigma(F^{-1}(\eta) - F^{-1}(\alpha))$.

Cas c : $H_1 : \theta < 2$**.** Dans ce cas le test est de la forme $\theta = 2 \Leftrightarrow \bar{X}_n > \kappa$ et puisque l'hypothèse alternative est composite, il faut résoudre le système

$$\left\{ \begin{array}{l} \alpha = \mathbb{P}(N(2, \sigma/\sqrt{n}) < \kappa) = F(\sqrt{n}(\kappa - 2)/\sigma) \\ \eta = \inf_{\theta < 2} \mathbb{P}(N(\theta, \sigma/\sqrt{n}) < \kappa) = \inf_{\theta < 2} F(\sqrt{n}(\kappa - \theta)/\sigma) = F(\sqrt{n}(\kappa - 2)/\sigma) \end{array} \right.$$

Donc dans ce cas, la puissance et égale au risque de première espèce ! Dans ce cas, les hypothèses nulle et alternative sont trop "proches", de telle sorte qu'on ne peut les distinguer avec certitude.

3. Par définition, dans le cas d'hypothèses simples le risque de première espèce est donné par

$$\alpha = \mathbb{P}_{\theta_0}(\text{Rejeter } H_0) = \mathbb{P}_{\theta_0}(\text{Rejeter } H_0) = \mathbb{P}_{\theta_0}\left(\left|\hat{\theta}_n - \theta_0\right| \geq \kappa\right)$$

et est donc décroissant en κ. A l'inverse, le risque de deuxième espèce est donné par

$$\beta = \mathbb{P}_{\theta_1}(\text{Accepter } H_0) = \mathbb{P}_{\theta_1}(\text{Accepter } H_1) = \mathbb{P}_{\theta_1}\left(\left|\hat{\theta}_n - \theta_0\right| < \kappa\right)$$

et est décroissant en κ. Ainsi, les risques de première et deuxième espèce varient en sens inverse. Fixer un risque de première espèce très faible présente donc un problème si l'on ne rejette pas H_0 : dans ce cas, le risque de deuxième espèce est élevé ce qui correspond au fait que si H_1 était vraie, alors la probabilité d'accepter H_0 serait élevée. Ainsi, avoir accepté H_0 n'est pas significatif.

↪ Exercice 6.2

1. Ces tests sont raisonnables car $\bar{X}_n$ ainsi que M_n ont "tendance" à être plus grands sous H_1 que sous H_0 : ainsi, plus ces statistiques seront petites et plus cela sera compatible avec H_0. Par ailleurs, sous H_0 on a $\bar{X}_n \to \frac{1}{2}$ et $M_n \to 1$, ce qui justifie de prendre $\kappa^X < 1/2$ et $\kappa^M < 1$, avec les valeurs précises calculées après.

2. Sous H_0, on a $\sqrt{n}(\bar{X}_n - 1/2)$ qui converge en loi vers une loi normale centrée et de variance $\sigma^2 = 1/12$ par le théorème central limite.

3. On a donc

$$\alpha = \mathbb{P}_1(\bar{X}_n > \kappa^X) \approx \mathbb{P}\left(\frac{\sigma}{\sqrt{n}} N(0,1) > \kappa^X - \frac{1}{2}\right) = 1 - F\left(\sqrt{n}(\kappa^X - 1/2)/\sigma\right)$$

ce qui donne la formule annoncée pour κ^X. Quant à κ^M, on a

$$\alpha = \mathbb{P}_1(M_n > \kappa^M) = 1 - (\kappa^M)^n$$

ce qui donne la formule annoncée pour κ^M.

4. Les puissances sont données par

$$\eta^X = \mathbb{P}_{1+\varepsilon}\left(\bar{X}_n > \kappa^X\right) \approx 1 - F\left(\frac{2\sqrt{3n}}{1+\varepsilon}\left(\kappa^X - \frac{1}{2} - \frac{\varepsilon}{2}\right)\right).$$

Pour le deuxième test, on a

$$\eta^M = \mathbb{P}_{1+\varepsilon}\left(M_n > \kappa^M\right) = \mathbb{P}_1\left((1+\varepsilon)M_n > \kappa^M\right) = 1 - \frac{(\kappa^M)^n}{(1+\varepsilon)^n} = 1 - \frac{1-\alpha}{(1+\varepsilon)^n}$$

5. Pour ε petit on a $-\ln(1-\eta^X) < -\ln(1-\eta^M)$ et donc la puissance du test basé sur M est plus grande.

6. La statistique de test dans le test de Neyman–Pearson est donnée par le rapport des vraisemblances : ici, on a $\mathcal{L}(\theta;\mathbf{x}) = \theta^{-n}\mathbb{1}\{0 \le x_1,\dots,x_n \le \theta\}$ et donc

$$\frac{\mathcal{L}(\theta_1;\mathbf{X}_n)}{\mathcal{L}(\theta_0;\mathbf{X}_n)} = \frac{\mathcal{L}(1+\varepsilon;\mathbf{X}_n)}{\mathcal{L}(1;\mathbf{X}_n)} = (1+\varepsilon)^{-n}\frac{\mathbb{1}\{M_n \le 1+\varepsilon\}}{\mathbb{1}\{M_n \le 1\}}.$$

Avec la convention $x/0 = +\infty$, on a donc

$$\frac{\mathcal{L}(\theta_1;\mathbf{X}_n)}{\mathcal{L}(\theta_0;\mathbf{X}_n)} = \begin{cases} (1+\varepsilon)^{-n} & \text{si } M_n \le 1, \\ +\infty & \text{sinon.} \end{cases}$$

Le test de Neyman–Pearson est de la forme

$$\frac{\mathcal{L}(\theta_1;\mathbf{X}_n)}{\mathcal{L}(\theta_0;\mathbf{X}_n)} > \kappa$$

et donc avec $\kappa > (1+\varepsilon)^{-n}$, on ne rejette H_0 que lorsque $M_n > 1$ auquel cas on ne se trompe jamais : le risque de première espèce est nul ! Si par contre $\kappa < (1+\varepsilon)^{-n}$ alors on rejette tout le temps H_0 si H_0 est vraie : on se trompe toujours ! Dans ce cas on ne peut donc fixer le risque de première espèce à une valeur $\alpha \in]0,1[$. Par ailleurs, la puissance du test vaut alors

$$\eta = \mathbb{P}_{1+\varepsilon}(1 < M_n < 1+\varepsilon) = 1 - \frac{1}{(1+\varepsilon)^n}$$

qui est du même ordre de grandeur que le test proposé à base de M_n.

↪ **Exercice 6.3**

1. Le test de Neyman–Pearson est de la forme

$$\text{Rejeter } H_0 \iff \frac{p_1(X,Y)}{p_0(X,Y)} > \kappa$$

où κ est déterminé par $\mathbb{P}_{H_0}(\text{Rejeter } H_0) = \alpha$. Utilisant les expressions de p_1 et p_0, on obtient

$$\text{Rejeter } H_0 \iff |X|,|Y| \le d, X^2+Y^2 > \kappa'$$

et κ' est défini par $\mathbb{P}_{H_0}(|X|,|Y| \le d, X^2+Y^2 > \kappa') = \alpha$.

2. On a

$$\mathbb{P}_{H_0}\left(|X|,|Y| \le d, X^2+Y^2 > G_2^{-1}(1-\alpha)\right) \le \mathbb{P}_{H_0}\left(X^2+Y^2 > G_2^{-1}(1-\alpha)\right)$$

qui vaut α puisque sous H_0, X et Y sont des variables normales centrées réduites indépendantes et donc X^2+Y^2 suit une loi du χ^2 à 2 degrés de liberté.

3. Pour $\alpha = 5\%$, la région critique est donc l'intersection entre l'extérieur du disque de rayon $\sqrt{6}$ et le carré $[-2,2]^2$, qui n'est pas vide puisque le coin du carré est de coordonnée $(\sqrt{8},\sqrt{8})$.
En revanche, pour $\alpha = 1\%$ l'intersection entre l'extérieur du disque de rayon $\sqrt{9} = 3$ et le carré $[-2,2]^2$ est vide. Cela signifie que dans ce cas, une seule observation ne permet pas de rejeter H_0 avec un risque de première espèce inférieure à 1%.

↪ **Exercice 6.4** *(Comparaison de moyenne et de variance)*

1. On sait d'après le cours que $(n_X-1)S_X^2/\sigma_X^2$ et $(n_Y-1)S_Y^2/\sigma_Y^2$ suivent des lois du χ^2 à n_X-1 et n_Y-1 degrés de liberté. Ainsi, $(S_X^2/\sigma_X^2)/(S_Y^2/\sigma_Y^2)$ suit une loi de Fisher–Snedecor de paramètre (n_X-1, n_Y-1).

Sous H_0, $\sigma_X = \sigma_Y$ et donc $S_X^2 = S_Y^2 = (S_X^2/\sigma_X^2)/(S_Y^2/\sigma_Y^2)$ suit une loi de Fisher–Snedecor de paramètre $(n_X - 1, n_Y - 1)$.

2. Sous l'hypothèse H_0, on a $S_X^2/S_Y^2 = (S_X^2/\sigma_X^2)/(S_Y^2/\sigma_Y^2)$ et S_X^2/S_Y^2 suit une loi de Fisher–Snedecor de paramètre $(n_X - 1, n_Y - 1)$ dont on notera F_{n_X-1,n_Y-1} la fonction de répartition. Pour $H_1 : \sigma_X > \sigma_Y$ on a donc le test suivant :

$$\text{Rejeter } H_0 \iff \frac{S_X^2}{S_Y^2} \geq \kappa$$

avec κ déterminé par

$$\mathbb{P}_{H_0}(\text{Rejeter } H_0) = \alpha \iff \kappa = F^{-1}_{n_X-1,n_Y-1}(1-\alpha).$$

Au final, le test de $H_0 : \sigma_X = \sigma_Y$ contre $H_1 : \sigma_X > \sigma_Y$ au niveau α est donné par

$$\text{Rejeter } H_0 \iff \frac{S_X^2}{S_Y^2} \geq F^{-1}_{n_X-1,n_Y-1}(1-\alpha).$$

3. Si l'on teste H_0 contre $H_1 : \sigma_X \neq \sigma_Y$, on considère alors un test bilatéral de la forme

$$\text{Rejeter } H_0 \iff \frac{S_X^2}{S_Y^2} \geq \kappa_+ \text{ ou } \frac{S_X^2}{S_Y^2} \leq \kappa_-$$

avec

$$\alpha = \mathbb{P}_{H_0}(\text{Rejeter } H_0) = \mathbb{P}_{H_0}\left(\frac{S_X^2}{S_Y^2} \geq \kappa_+ \text{ ou } \frac{S_X^2}{S_Y^2} \leq \kappa_-\right) = 1 - F_{n_X-1,n_Y-1}(\kappa_+) + F_{n_X-1,n_Y-1}(\kappa_-).$$

Si l'on impose l'hypothèse de symétrie $1-F_{n_X-1,n_Y-1}(\kappa_+) = F_{n_X-1,n_Y-1}(\kappa_-)$, on obtient $\kappa_+ = F^{-1}_{n_X-1,n_Y-1}(1-\alpha/2)$. On remarque donc que la forme du test est dictée par H_1.

4. On a $(n-2)S^2/\sigma^2 = \sum_{k=1}^{n_X}(X_k - \bar{X})^2/\sigma^2 + \sum_{k=1}^{n_Y}(Y_k - \bar{Y})^2/\sigma^2$. Chaque somme suit une loi du χ^2 à $n_X - 1$ et $n_Y - 1$ degrés de liberté, et comme les X_i et les Y_i sont indépendants la somme totale et donc $(n-2)S^2/\sigma^2$ suit une loi du χ^2 à $n-2$ degrés de liberté.
Par ailleurs, $\bar{X} - \bar{Y}$ suit une loi normale de moyenne $m_X - m_Y$ et de variance $\sigma_X^2/n_X + \sigma_Y^2/n_Y = \sigma^2(1/n_X + 1/n_Y)$.

5. On sait d'après le cours sur les vecteurs gaussiens que S_X et $\bar{X}$ sont indépendants, ainsi que S_Y et $\bar{Y}$. Il s'ensuit donc bien que S est indépendante de $\bar{X}$ et $\bar{Y}$. Ainsi, $(\bar{X} - \bar{Y} - m_X + m_Y)/(\sigma\sqrt{1/n_X + 1/n_Y})$ suit une loi normale centrée réduite et est indépendante de $(n-2)S^2/\sigma^2$ qui suit une loi du χ^2 à $n-2$ degrés de liberté : ainsi,

$$\frac{(\bar{X} - \bar{Y} - m_X + m_Y)/(\sigma\sqrt{1/n_X + 1/n_Y})}{\sqrt{((n-2)S^2/\sigma^2)/(n-2)}} = \frac{\bar{X} - \bar{Y} - m_X + m_Y}{S\sqrt{1/n_X + 1/n_Y}}$$

suit une loi de Student à $n-2$ degrés de liberté.

6. On considère le test

$$\text{Rejeter } H_0 \iff \frac{|\bar{X} - \bar{Y}|}{S\sqrt{1/n_X + 1/n_Y}} \geq T_{n-2}^{-1}(1-\alpha/2)$$

avec T_{n-2} la fonction de répartition de la loi de Student à $n-2$ degrés de liberté. Ainsi, on a prouvé dans la question précédente que pour ce test,

$$\begin{aligned}
\mathbb{P}_{H_0}(\text{Rejeter } H_0) &= \mathbb{P}_{H_0}\left(\frac{|\bar{X} - \bar{Y}|}{S\sqrt{1/n_X + 1/n_Y}} \geq T_{n-2}^{-1}(1-\alpha/2)\right) \\
&= 1 - T_{n-2}(T_{n-2}^{-1}(1-\alpha/2)) + T_{n-2}(-T_{n-2}^{-1}(1-\alpha/2)) \\
&= 1 - (1-\alpha/2) + T_{n-2}(T_{n-2}^{-1}(1-\alpha/2)) \\
&= \alpha/2 + \alpha/2 \\
&= \alpha
\end{aligned}$$

et on contrôle donc bien le risque de première espèce. Pour la troisième égalité, on a utilisé la relation $T_{n-2}^{-1}(x) = -T_{n-2}^{-1}(1-x)$. En effet

$$\begin{aligned} T_{n-2}^{-1}(x) = -T_{n-2}^{-1}(1-x) &\Longleftrightarrow x = T_{n-2}\left(-T_{n-2}^{-1}(1-x)\right) \\ &\Longleftrightarrow x = 1 - T_{n-2}\left(T_{n-2}^{-1}(1-x)\right) \end{aligned}$$

où la deuxième équivalence provient de la symétrie de T_{n-2} : pour tout x, $T_{n-2}(x) + T_{n-2}(-x) = 1$.

↪ Problème 6.5

1. Les variables aléatoires sont indépendantes et de même loi $\mathcal{N}(\mu, \sigma^2)$, avec μ inconnu.

2. L'estimateur de la moyenne empirique $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$ est un estimateur sans biais de μ. Dans ce modèle, c'est également l'estimateur du maximum de vraisemblance. La loi de $\hat{\mu}_n$ est la loi gaussienne $\mathcal{N}(\mu, \sigma^2/n)$.

3. On considère l'hypothèse nulle $H_0 : \mu \leq \mu_0$ (le nouveau modèle n'est pas plus performant que l'ancien) et l'hypothèse alternative $H_1 : \mu > \mu_0$ ce qui donne un test de la forme

$$\text{Rejeter } H_0 \Longleftrightarrow \hat{\mu}_n \geq \kappa$$

avec κ qui détermine le risque de première espèce α. Puisque H_0 est composite, on a

$$\alpha = \sup_{\mu \in H_0} \mathbb{P}_\mu(\hat{\mu}_n \geq \kappa) = \sup_{\mu \leq \mu_0} \mathbb{P}_\mu(\hat{\mu}_n \geq \kappa).$$

Comme $\hat{\mu}_n$ est de loi gaussienne $\mathcal{N}(\mu, \sigma^2/n)$, $\sqrt{n}(\hat{\mu}_n - \mu)/\sigma$ est centre réduit et il vient

$$\mathbb{P}_\mu(\hat{\mu}_n \geq \kappa) = 1 - F\left(\sqrt{n}\frac{\kappa - \mu}{\sigma}\right)$$

avec F la fonction de répartition de $\mathcal{N}(0,1)$. On remarque que $\mu \mapsto \mathbb{P}_\mu(\hat{\mu}_n \geq \kappa)$ est une fonction croissante de μ et donc $\sup_{\mu \leq \mu_0} \mathbb{P}_\mu(\hat{\mu}_n \geq \kappa) = \mathbb{P}_{\mu_0}(\hat{\mu}_n \geq \kappa)$. Pour un seuil α, κ est donc déterminé par l'équation :

$$\alpha = 1 - F\left(\sqrt{n}\frac{\kappa - \mu_0}{\sigma}\right)$$

soit $\kappa = \mu_0 + \sigma F^{-1}(1-\alpha)/\sqrt{n}$. Pour l'application numérique, on a $\alpha = 5\%$, $F^{-1}(1-\alpha) \approx 1.64$ et $\kappa \approx 123.9 > \hat{\mu}_n = 123.5$.

4. On évalue l'erreur de deuxième espèce, β, pour $\mu_1 = 1.05\mu_0$, c'est-à-dire la probabilité de rejeter le nouveau modèle sachant que l'annonce du représentant est exacte et donc que le nouveau modèle est plus performant de 5%. Elle est définie par :

$$\beta = \mathbb{P}_{\mu_1}\left(\hat{\mu}_n < \kappa\right) = \mathbb{P}_{\mu_1}\left(\frac{\sqrt{n}(\hat{\mu}_n - \mu_1)}{\sigma} < \frac{\sqrt{n}(\kappa - \mu_1)}{\sigma}\right).$$

Sous $\mathbb{P}_{\mu_1}$, $\hat{\mu}_n \sim \mathcal{N}(\mu_1, \sigma^2/n)$ et donc, poursuivant les calculs précédents, on obtient

$$\beta = F\left(\frac{\sqrt{n}(\kappa - \mu_1)}{\sigma}\right) = F\left(F^{-1}(1-\alpha) + \frac{\sqrt{n}(\mu_0 - \mu_1)}{\sigma}\right).$$

L'application numérique donne $\beta \approx 19.5\%$. Il s'agit du risque du vendeur, i.e. le risque que sa machine ne soit pas achetée alors qu'elle est 5% meilleure que l'ancienne.

5. L'hypothèse nulle est alors $H_0' : \mu = \mu_1$, et l'hypothèse alternative $H_1 : \mu \leq \mu_0$. Dans ce cas, le test est de la forme

$$\text{Rejeter } H_0 \Longleftrightarrow \hat{\mu}_n \leq \kappa'$$

et l'erreur de première espèce est :

$$\mathbb{P}_{\mu_1}(\hat{\mu}_n \leq \kappa') = F\left(\sqrt{n}\frac{\kappa' - \mu_1}{\sigma}\right).$$

Pour un niveau α, on obtient :

$$\kappa' = \mu_1 + \frac{1}{\sqrt{n}}\sigma F^{-1}(\alpha) = 1.05\mu_0 + \frac{1}{\sqrt{n}}\sigma F^{-1}(\alpha)$$

Pour $\alpha = 5\%$, on obtient $\kappa' \approx 122.0 < \hat{\mu}_n = 123.5$. On accepte donc H_0' : le nouveau modèle est donc plus performant que l'ancien de 5% en moyenne. Le risque de deuxième espèce maximal est alors ;

$$\begin{aligned}
\beta' &= \sup_{\mu \leq \mu_0} \mathbb{P}_\mu(\hat{\mu}_n > \kappa') \\
&= \mathbb{P}_{\mu_0}(\hat{\mu}_n > \kappa') \\
&= 1 - F\left(F^{-1}(\alpha) - \sqrt{n}\frac{\mu_0 - \mu_1}{\sigma}\right) \\
&= F\left(-F^{-1}(\alpha) + \sqrt{n}\frac{\mu_0 - \mu_1}{\sigma}\right) \\
&= \beta.
\end{aligned}$$

La quatrième égalité vient de $1 - F(x) = F(-x)$ et la dernière de $-F^{-1}(x) = F^{-1}(1-x)$ (cf. correction de la dernière question de l'exercice 6.3) et de l'expression de β calculée précédemment. L'acheteur a donc environ 20% de chance d'accepter l'annonce du représentant alors que celle-ci est fausse.

6. **A FAIRE**

Annexe B

Tableau des lois usuelles

Les pages suivantes présentent la définition des lois discrètes et absolument continues les plus usuelles, avec leur domaine de définition et leurs transformées.

Loi	Paramètre	Support	Probabilité élémentaire $\mathbb{P}(X = k)$	Fonction de répartition $F_X(x)$	Espérance	Variance	Fonction caractéristique $\varphi_X(t)$
Uniforme	$n \in \mathbb{N}^*$	$\{1, \dots, n\}$	$\frac{1}{n}$	$\frac{x}{n}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	$\frac{1-e^{int}}{n(e^{-it}-1)}$
Bernoulli	$p \in]0,1[$	$\{0,1\}$	$\begin{cases} p & \text{si } k=1 \\ 1-p & \text{si } k=0 \end{cases}$	$\begin{cases} 1-p & \text{si } x=0 \\ 1 & \text{si } x=1 \end{cases}$	p	$p(1-p)$	$pe^{it}+1-p$
Binomiale	$n \in \mathbb{N}^*$, $p \in]0,1[$	$\{0, \dots, n\}$	$\binom{n}{k} p^k (1-p)^{n-k}$	$\sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k}$	np	$np(1-p)$	$\left(pe^{it}+1-p\right)^n$
Géométrique	$p \in]0,1[$	$\mathbb{N}^*$	$(1-p)^{k-1}p$	$1-(1-p)^x$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^{it}}{1-e^{it}+pe^{it}}$
Poisson	$\mu > 0$	$\mathbb{N}$	$e^{-\mu}\frac{\mu^k}{k!}$	$\sum_{k=0}^{x} e^{-\mu}\frac{\mu^k}{k!}$	μ	μ	$\exp\left(-\mu(1-e^{it})\right)$

TABLEAU B.1 – Lois discrètes classiques : dans les colonnes probabilité élémentaire et fonction de répartition, on ne considère que k et x dans le support. De manière plus générale, la loi uniforme est définie sur n'importe quel ensemble fini A et est définie par $\mathbb{P}(X = a) = \frac{1}{|A|}$. Puisque toutes les lois présentées sont des lois sur $\mathbb{N}$, leur transformée de Laplace L_X et fonction génératrice φ_X sont aussi définies. Elles s'obtiennent directement à partir de la fonction caractéristique : $L_X(\lambda) = \varphi_X(i\lambda)$ et $\phi_X(z) = L_X(-\ln z)$.

Loi	Paramètre	Support	Densité $f_X(x)$	Fonction de répartition $F_X(x)$	Espérance	Variance	Fonction caractéristique $\varphi_X(t)$
Uniforme (continue)	$a < b$	$[a, b]$	$\dfrac{1}{b-a}$	$\dfrac{x-a}{b-a}$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$	$\dfrac{e^{itb} - e^{ita}}{it(b-a)}$
Normale	$m \in \mathbb{R}$ $\sigma > 0$	$\mathbb{R}$	$\dfrac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\dfrac{(x-m)^2}{2\sigma^2}\right)$	$\displaystyle\int_{-\infty}^{x} f_X(y)\mathrm{d}y$	m	σ^2	$\exp\left(-\dfrac{1}{2}t^2\sigma^2 + itm\right)$
Exponentielle	$\lambda >$	$[0, \infty[$	$\lambda e^{-\lambda x}$	$1 - e^{-\lambda x}$	$\dfrac{1}{\lambda}$	$\dfrac{1}{\lambda}$	$\dfrac{\lambda}{\lambda - it}$
Gamma	$\alpha, \beta > 0$	$[0, \infty[$	$\dfrac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$	$\displaystyle\int_0^x f_X(y)\mathrm{d}y$	$\dfrac{\alpha}{\beta}$	$\dfrac{\alpha}{\beta^2}$	$\left(1 - \dfrac{it}{\beta}\right)^{-\alpha}$

TABLEAU B.2 – Lois absolument continues classiques : dans les colonnes densité et fonction de répartition, on ne considère que x dans le support.