Probabilités et Statistique 2

Pierre ÉTORÉ
Olivier GAUDOIN

Ensimag - première année

Année 2025-2026

Table des matières

Chapitre 1

Espaces probabilisés

Dans ce chapitre on introduit les notions de base de probabilités, qui aboutissent au passage à la notion d'espace de probabilité (ou espace probabilisé). Puis on explore certaines premières propriétés de ces espaces. On s'appuie beaucoup (dans ce chapitre mais aussi dans des chapitres ultérieurs) sur des notions d'intégration et de théorie de la mesure, et on se réfèrera le cas échéant au polycopié d'Analyse pour l'ingénieur avancée (sem. 6). Cependant dans un souci de cohésion du présent polycopié il nous arrivera aussi de redéfinir certaines notions d'intégration.

1.1 Tribus, espaces mesurables et mesures

Pour E ensemble on notera de façon standard $\mathcal{P}(E)$ l'ensemble des parties (ou sous-ensembles) de E. Pour $A \subset E$ (i.e. $A \in \mathcal{P}(E)$) on note $A^c = E \setminus A$ le complémentaire de A dans E.

Voici une première définition.

Définition 1.1.1. Soit E un ensemble et $\mathcal{A}$ un ensemble de parties de E (i.e. $\mathcal{A} \subset \mathcal{P}(E)$). On dit que $\mathcal{A}$ est une algèbre (de parties de E) si
i) On a $E \in \mathcal{A}$.
ii) Pour tout $A \in \mathcal{A}$ on a $A^c \in \mathcal{A}$.
iii) Pour toute suite finie $(A_i)_{i=0}^n$ d'éléments de $\mathcal{A}$ (ici $n \in \mathbb{N}$) on a $\bigcup_{i=0}^n A_i \in \mathcal{A}$.

Si dans la définition ci-dessus on remplace le point iii) par une clause de stabilité par réunion infinie dénombrable on obtient la nouvelle définition que voici :

Définition 1.1.2. Soit E un ensemble et $\mathcal{E}$ un ensemble de parties de E (i.e. $\mathcal{E} \subset \mathcal{P}(E)$). On dit que $\mathcal{E}$ est une tribu (de parties de E, on dit aussi *tribu sur E*) si
i) On a $E \in \mathcal{E}$.
ii) Pour tout $A \in \mathcal{E}$ on a $A^c \in \mathcal{E}$.
iii) Pour toute suite $(A_i)_{i\in\mathbb{N}}$ d'éléments de $\mathcal{E}$ on a $\bigcup_{i\in\mathbb{N}} A_i \in \mathcal{E}$.

Remarque 1.1.1. Soit $\mathcal{E}$ une tribu sur E. Notons que par passage au complémentaire on a toujours $\emptyset \in \mathcal{E}$ et que si (A_i) est une suite d'éléments de $\mathcal{E}$ on a toujours $\bigcap_i A_i \in \mathcal{E}$ (respectivement car $E^c = \emptyset$ et $\bigcap_i A_i = \left(\bigcup_i A_i^c\right)^c$).

Remarque 1.1.2. En fait dans la définition 1.1.2 on peut toujours remplacer de façon équivalente i) par
i') On a $\emptyset \in \mathcal{E}$
et iii) par
iii') Pour toute suite $(A_i)_{i\in\mathbb{N}}$ d'éléments de $\mathcal{E}$ on a $\bigcap_{i\in\mathbb{N}} A_i \in \mathcal{E}$.

Il est clair qu'une tribu $\mathcal{E}$ sur E est a fortiori une algèbre de parties de E (pour $n \in \mathbb{N}$ et $(A_i)_{i=0}^n$ suite finie d'éléments de $\mathcal{E}$ il suffit de considérer la suite $(B_i)_{i\in\mathbb{N}}$ définie par $B_i = A_i$ pour $i = 0, \ldots, n$, et $B_i = \emptyset$ pour $i \geq n+1$ et d'appliquer le point iii) de la définition 1.1.2 ; le point iii) de la définition 1.1.1 est alors vérifié).

Exemple 1.1.1. 1) La tribu grossière $\mathcal{E} = \{E, \emptyset\}$; c'est la tribu la moins fine qu'on peut mettre sur E.

2) La tribu la plus fine qu'on peut mettre sur E c'est $\mathcal{P}(E)$ (il est bien clair que c'est une tribu ; attention au vocabulaire c'est la plus "grosse" au sens de l'inclusion).

3) Soit $\mathcal{C} \subset \mathcal{P}(E)$, l'ensemble $\mathcal{C}$ n'étant pas forcément une tribu. On note $\sigma(\mathcal{C})$ la tribu engendrée par $\mathcal{C}$, i.e. la plus petite tribu qui contient $\mathcal{C}$ (en ce sens que si $\mathcal{Y}$ est une tribu t.q. $\mathcal{C} \subset \mathcal{Y}$ alors $\sigma(\mathcal{C}) \subset \mathcal{Y}$). Une telle tribu existe toujours (cf définition 2.2 du cours Analyse pour l'ingénieur avancée).

4) La tribu des boréliens sur $\mathbb{R}^d$, notée $\mathcal{B}(\mathbb{R}^d)$. C'est la tribu engendrée par les ouverts de $\mathbb{R}^d$, au sens du point 3) ci-dessus (noter que l'ensemble des ouverts de $\mathbb{R}^d$ n'est ni une tribu ni une algèbre sur $\mathbb{R}^d$; en effet cet ensemble n'est pas stable par passage au complémentaire puisque les fermés de $\mathbb{R}^d$ n'y sont pas).

Noter que $\mathcal{B}(\mathbb{R}^d)$ est strictement plus grosse que l'ensemble des ouverts de $\mathbb{R}^d$ mais strictement plus petite que $\mathcal{P}(\mathbb{R}^d)$ (cf par exemple cf [2], Chap.2).

Nous rencontrerons fréquemment dans la suite de ce polycopié la tribu des boréliens, et aurons l'occasion d'illustrer son rôle fondamental.

Nous voyons que l'inclusion fournit une relation d'ordre sur les tribus. Si pour $\mathcal{E}, \mathcal{G}$ tribus sur E on a $\mathcal{G} \subset \mathcal{E}$ on dit que $\mathcal{G}$ est *sous-tribu* de $\mathcal{E}$.

Un ensemble E muni d'une tribu $\mathcal{E}$ sur E est appelé *espace mesurable*. On note $(E, \mathcal{E})$ une telle entité. Cette terminologie signifie que $(E, \mathcal{E})$ est prêt à être équipé d'une *mesure*, dont voici la définition :

Définition 1.1.3. Soit $(E, \mathcal{E})$ un espace mesurable.

Une mesure sur $(E, \mathcal{E})$ est une application $\mu : \mathcal{E} \to [0, +\infty]$ qui vérifie $\mu(\emptyset) = 0$ et est σ-additive, i.e. si $(A_i)_{i \in \mathbb{N}}$ est une suite d'éléments disjoints de $\mathcal{E}$ alors $\mu\big(\cup_{i=0}^{\infty} A_i\big) = \sum_{i=0}^{\infty} \mu(A_i)$.

La définition ci-dessus correspond plus précisément à celle d'une mesure positive (par opposition à signée). Les mesures rencontrées dans ce polycopié seront toutes positives, nous dirons donc simplement "mesure".

Une mesure sur $(E, \mathcal{E})$ attribue donc une masse à des parties de E qui sont dans la tribu $\mathcal{E}$. Il se peut que cette masse soit infinie (sauf si on a affaire à une *mesure finie* comme par exemple dans le point ii) de la proposition 1.1.1 ci-après).

Si μ est une mesure sur $(E, \mathcal{E})$ espace mesurable on appelle le triplet $(E, \mathcal{E}, \mu)$ un *espace mesuré*.

Voici quelques premières propriétés faciles à établir d'une mesure :

Propriété 1.1.1. Soit $(E, \mathcal{E}, \mu)$ un espace mesuré.

i) La mesure μ est *additive* en ce sens que pour toute suite finie $(A_i)_{i=0}^n$, $n \in \mathbb{N}$, d'éléments disjoints de $\mathcal{E}$ on a $\mu\big(\cup_{i=0}^{n} A_i\big) = \sum_{i=0}^{n} \mu(A_i)$.

ii) La mesure μ est croissante en ce sens que $\mu(A) \leq \mu(B)$ pour tous $A, B \in \mathcal{E}$ avec $A \subset B$.

Démonstration. i) A nouveau pour $n \in \mathbb{N}$ et $(A_i)_{i=0}^n$ suite finie d'éléments disjoints de $\mathcal{E}$ il suffit de considérer la suite $(B_i)_{i \in \mathbb{N}}$ définie par $B_i = A_i$ pour $i = 0, \ldots, n$, et $B_i = \emptyset$ pour $i \geq n+1$. En appliquant la relation de σ-additivité $\mu\big(\cup_i B_i\big) = \sum_i \mu(B_i)$ la relation d'additivité $\mu\big(\cup_{i=0}^{n} A_i\big) = \sum_{i=0}^{n} \mu(A_i)$ vient naturellement.

ii) Il suffit de remarquer que si $A \subset B$ alors $B = A \cup (B \setminus A)$ cette réunion étant disjointe (remarquons au passage que $B \setminus A = B \cap A^c$ est bien dans $\mathcal{E}$). On a donc par additivité

$$\mu(B) = \mu(A) + \mu(B \setminus A),$$

ce qui amène la relation voulue puisque $\mu(B \setminus A) \geq 0$ (mesure positive). □

Les propriétés suivantes sont moins immédiates, nous les énonçons dans une proposition (similaire à la proposition 2.1 du cours Analyse pour l'ingénieur avancée) :

Proposition 1.1.1. *Soit $(E, \mathcal{E}, \mu)$ un espace mesuré.*

i) On a la propriété de continuité séquentielle croissante de la mesure : pour toute suite croissante (A_i) d'éléments de $\mathcal{E}$ (i.e. $A_i \subset A_{i+1}$ pour tout i), $\lim_i \mu(A_i) = \mu(\cup_i A_i)$.

ii) Si $\mu(E) < \infty$ (on dit alors que μ est une mesure finie) on a aussi la propriété de continuité séquentielle décroissante de la mesure μ : pour toute suite décroissante (A_i) d'éléments de $\mathcal{E}$ (i.e. $A_{i+1} \subset A_i$ pour tout i), $\lim_i \mu(A_i) = \mu(\cap_i A_i)$.

iii) La mesure μ est σ-sous-additive : i.e. pour (A_i) suite quelconque d'éléments de $\mathcal{E}$ (non forcément disjoints) on a $\mu(\cup_i A_i) \leq \sum_i \mu(A_i)$.

Démonstration. i) On définit $B_i = A_{i+1} \setminus A_i$ pour tout $i \in \mathbb{N}$. Comme la suite (A_i) est croissante on a que les B_i sont disjoints, que $A_i = A_0 \cup \bigcup_{0 \leq k \leq i-1} B_k$ pour tout $i \geq 1$, et que $\bigcup_i A_i = A_0 \cup \bigcup_{i \geq 0} B_i$. Par σ-additivité de μ on a donc

$$\begin{aligned}
\mu(\bigcup_i A_i) &= \mu(A_0) + \sum_{k \geq 0} \mu(B_k) \\
&= \mu(A_0) + \lim_{i \to \infty} \sum_{0 \leq k \leq i-1} \mu(B_k) \\
&= \lim_{i \to \infty} \Big(\mu(A_0) + \sum_{0 \leq k \leq i-1} \mu(B_k)\Big) \\
&= \lim_{i \to \infty} \mu(A_i).
\end{aligned}$$

ii) Comme la suite (A_i) est décroissante on remarque que la suite (A_i^c) est croissante. De plus pour tout $B \in \mathcal{E}$ on a par additivité $\mu(B) + \mu(B^c) = \mu(E)$ et donc la relation $\mu(B) = \mu(E) - \mu(B^c)$ (qui fait sens puisque $\mu(E) < \infty$). On peut donc utiliser cette dernière relation et le point i) qui amènent

$$\mu(\bigcap_i A_i) = \mu(E) - \mu(\bigcup_i A_i^c) = \mu(E) - \lim_i \mu(A_i^c) = \lim_i (\mu(E) - \mu(A_i^c)) = \lim_i \mu(A_i),$$

qui est le résultat voulu.

iii) On traite d'abord le cas d'un ensemble fini d'indices $I = \{0, 1, \ldots, n\}$. En remarquant que

$$\mu(A_0 \cup A_1) = \mu(A_0 \cup (A_1 \setminus A_0)) = \mu(A_0) + \mu(A_1 \setminus A_0) \leq \mu(A_0) + \mu(A_1)$$

et en généralisant ce type de relation on peut montrer par récurrence que $\mu(\cup_{i \in I} A_i) \leq \sum_{i \in I} \mu(A_i)$.

On a alors pour tout $n \in \mathbb{N}$,

$$\mu(\bigcup_{0 \leq i \leq n} A_i) \leq \sum_{0 \leq i \leq n} \mu(A_i) \leq \sum_{i \in \mathbb{N}} \mu(A_i).$$

Mais il est clair que la suite $\big(\bigcup_{0 \leq i \leq n} A_i\big)_n$ est une suite croissante d'évènements donc par le point i) on a

$$\lim_n \mu(\bigcup_{0 \leq i \leq n} A_i) = \mu(\bigcup_{n \in \mathbb{N}} \bigcup_{0 \leq i \leq n} A_i) = \mu(\bigcup_{i \in \mathbb{N}} A_i).$$

Comme le passage à la limite conserve les inégalités larges on a bien montré la σ-sous-additivité. (on renvoie à [1] proposition I.4.3 pour plus de détails). □

On donne ici la définition de propriété vraie μ-presque partout, similaire à la définition 2.5 du cours Analyse pour l'ingénieur avancée :

Définition 1.1.4. Soit $(E, \mathcal{E}, \mu)$ un espace mesuré. Soit $P(x)$ une propriété dépendant de $x \in E$. On dit que $P(x)$ est vraie (μ-)presque partout (p.p.), ou pour (μ-)presque tout x, si l'ensemble $\{x \in E : P(x)$ n'est pas vraie$\}$ est (μ-)négligeable, i.e. inclus dans $B \in \mathcal{E}$ avec $\mu(B) = 0$.

Comment sont construites les mesures ?

Si E est dénombrable les choses sont assez simples et on peut utiliser le résultat suivant.

Proposition 1.1.2. *Soit E un espace fini ou infini dénombrable et considérons $\mathcal{E} = \mathcal{P}(E)$.*

i) Une mesure sur $(E, \mathcal{E})$ est caractérisée par ses valeurs sur les singletons : $\mu_x = \mu(\{x\})$, $x \in E$.

ii) Si $(\mu_x)_{x\in E}$ est une famille de réels positifs il existe une mesure μ sur $(E, \mathcal{E})$ (nécessairement unique par i)) vérifiant $\mu(\{x\}) = \mu_x$ pour tout $x \in E$.

Démonstration. Soit $A \in \mathcal{E}$ on a $A = \bigcup_{x\in A}\{x\}$ réunion disjointe dénombrable de singletons. Comme μ est σ-additive sur $\mathcal{E}$ on a

$$\mu(A) = \mu\Big(\bigcup_{x\in A}\{x\}\Big) = \sum_{x\in A}\mu(\{x\}) = \sum_{x\in A}\mu_x$$

ce qui montre i).

Soit $(\mu_x)_{x\in E}$ est une famille de réels positif. Pour $A \in \mathcal{E}$ on pose $\mu(A) := \sum_{x\in A}\mu_x$ avec la convention qu'une somme vide se voit attribuée la masse zéro. L'application μ ainsi construite est bien σ-additive car pour (A_i) suite d'éléments disjoints de $\mathcal{P}(E)$ on a

$$\mu\Big(\bigcup_i A_i\Big) = \sum_{x\in\cup_i A_i}\mu_x = \sum_i\sum_{x\in A_i}\mu_x = \sum_i \mu(A_i)$$

(on peut sommer par paquets). En conclusion on a bien défini une mesure sur $(E, \mathcal{E})$, ce qui prouve ii). □

Si E est infini non dénombrable la question peut être très délicate.

Prenons l'exemple bien connu de la mesure de Lebesgue, notée λ, sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Cette mesure est construite de telle sorte que pour tout intervalle I dans $\mathcal{B}(\mathbb{R})$ (c'est à dire $I = [a, b]$, $I = [a, b)$, $I = (a, b)$ ou $I = (a, b]$) on a $\lambda(I) = |I|$ (où $|I| = b - a$ est la "longueur" de I).

La mesure de Lebesgue semble donc naturelle pourtant sa construction est difficile. En effet la tribu des boréliens $\mathcal{B}(\mathbb{R})$ contient bien sûr les intervalles, mais aussi toute sorte de parties de $\mathbb{R}$ qui ne sont pas des intervalles. Comment alors définir l'application de λ à ce type de partie ?

Si on veut construire la mesure de Lebesgue sur l'intervalle $[0, 1]$ par exemple on peut utiliser le résultat suivant, connu sous le nom de théorème de Carathéodory.

Théorème 1.1.1. *Soit E un ensemble et $\mathcal{A}$ une algèbre de parties de E. Soit $\tilde{\mu} : \mathcal{A} \to [0, \infty]$ une application additive et finie (i.e. telle que $\tilde{\mu}(E) < \infty$). Alors il existe une unique mesure μ sur $(E, \sigma(\mathcal{A}))$ telle que pour tout $A \in \mathcal{A}$ on a $\mu(A) = \tilde{\mu}(A)$ (on dit que μ est une extension de $\tilde{\mu}$).*

Démonstration. Cf par exemple [5] paragraphe I.5. □

La démarche est alors la suivante : on considère l'ensemble $\mathcal{A}$ des réunions finies d'intervalles inclus dans $[0, 1]$. C'est une algèbre. On considère $\tilde{\lambda} : I \mapsto |I|$ pour $I \subset [0, 1]$ intervalle, qu'on prolonge par additivité à tous les éléments de $\mathcal{A}$ (i.e. on exige que si $\cup_{i=0}^n I_i \in \mathcal{A}$ est une réunion disjointes d'intervalles on a $\tilde{\lambda}(\cup_{i=0}^n I_i) = \sum_{i=0}^n \tilde{\lambda}(I_i)$). L'application $\tilde{\lambda}$ est donc par construction additive et on a $\tilde{\lambda}([0, 1]) = 1 < \infty$.

Par le théorème de Carathéodory $\tilde{\lambda}$ s'étend en une mesure λ sur $[0, 1]$ muni de la tribu $\sigma(\mathcal{A})$. Or il est possible de montrer que $\sigma(\mathcal{A})$ n'est autre que la tribu des boréliens sur $[0, 1]$, notée $\mathcal{B}([0, 1])$. Finalement λ est la mesure de Lebesgue sur $([0, 1], \mathcal{B}([0, 1]))$.

Evidemment pour construire la mesure de Lebesgue sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ il faut travailler plus puisque la longueur de $\mathbb{R}$ est infinie ! (noter que dans le cours <u>Analyse pour l'ingénieur avancée</u> l'existence de la mesure de Lebesgue sur $\mathbb{R}^d$ fait l'objet du théorème 2.3).

Remarque 1.1.3. Noter que la construction de la mesure de Lebesgue aboutit à une mesure sur $\mathbb{R}^d$ équipé de la tribu des boréliens $\mathcal{B}(\mathbb{R}^d)$. Il est impossible de construire une mesure avec les mêmes propriétés (en particulier que la mesure d'une portion de $\mathbb{R}^d$ est son volume) sur $\mathbb{R}^d$ équipé de $\mathcal{P}(\mathbb{R}^d)$: la tribu $\mathcal{P}(\mathbb{R}^d)$ est trop fine (ou trop grosse) pour permettre la construction de la mesure de Lebesgue.

Le théorème 1.1.1 sert aussi parfois à construire des mesures de probabilités, qui sont celles qui vont nous intéresser dans la section qui suit (cf le "long aparté" au sein de l'exemple 1.2.1). Il nous servira aussi plus loin à montrer que les fonctions de répartitions caractérisent complètement les lois de variables aléatoires (mais nous anticipons là sur le Chapitre 2, théorème 2.3.1).

1.2 Mesure de probabilité, espace de probabilité

Définition 1.2.1. Soit $(\Omega, \mathcal{F})$ un espace mesurable. Une mesure $\mathbb{P}$ sur $(\Omega, \mathcal{F})$ vérifiant $\mathbb{P}(\Omega) = 1$ est appelé mesure de probabilités. Le triplet $(\Omega, \mathcal{F}, \mathbb{P})$ est appelé espace de probabilité ou espace probabilisé.

Remarque 1.2.1. Le couple $(\Omega, \mathcal{F})$ est parfois appelé espace probabilisable.

Remarque 1.2.2. Dans certains ouvrages (par exemple [4]) on cherche à donner une définition minimaliste d'une mesure de probabilité sur $(\Omega, \mathcal{F})$. On dit : c'est une application $\mathbb{P} : \mathcal{F} \to [0, 1]$ qui vérifie
i) $\mathbb{P}(\Omega) = 1$
ii) $\mathbb{P}$ est σ-additive.

Puis on peut vérifier que $\mathbb{P}(\emptyset) = 0$ de diverses façons. Par exemple comme $\mathbb{P}$ est σ-additive elle est additive et donc $\mathbb{P}(\emptyset \cup \Omega) = \mathbb{P}(\emptyset) + \mathbb{P}(\Omega) = \mathbb{P}(\emptyset) + 1$. Mais bien sûr $\mathbb{P}(\emptyset \cup \Omega) = \mathbb{P}(\Omega) = 1$, d'où $\mathbb{P}(\emptyset) = 1 - 1 = 0$.

A l'arrivée la situation est le même : $\mathbb{P}$ est une mesure sur $(\Omega, \mathcal{F})$ vérifiant $\mathbb{P}(\Omega) = 1$.

Remarquons qu'une mesure de probabilité $\mathbb{P}$ est en particulier une mesure finie sur $(\Omega, \mathcal{F})$. Un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ récupère donc automatiquement toutes les propriétés des espaces mesurés avec mesure finie.

Commentons les objets en jeu dans la définition 1.2.1 :

L'ensemble Ω est l'*univers des possibles*, il contient des *états du monde* notés en général ω, correspondant aux issues possibles d'une expérience aléatoire.

La tribu $\mathcal{F}$ contient des parties $A \subset \Omega$ appelées *évènements*. Ce sont des regroupements d'états du monde correspondant à une situation donnée.

La mesure de probabilités $\mathbb{P}$ attribue un poids entre 0 et 1 aux évènements, en allant du moins probable au plus probable.

Prenons un exemple assez simple. On réalise une expérience aléatoire qui consiste à tirer un dé à six faces. Les issues possibles sont 1,2,3,4,5,6 on a donc $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Imaginons qu'on a $\mathcal{F} = \mathcal{P}(\Omega)$ et que le dé est équilibré. On a donc

$$\mathbb{P}(\{i\}) = 1/6 \quad \text{pour tout} \quad i = 1, \ldots, 6 \tag{1.2.1}$$

(noter que les singletons $\{i\}$ sont bien dans $\mathcal{F}$; de plus (1.2.1) suffit à définir $\mathbb{P}$ en tant que mesure de probabilité sur $(\Omega, \mathcal{F})$, par la proposition 1.1.2, et en remarquant que $\sum_{\omega \in \Omega} \mathbb{P}(\{\omega\}) = 1$).

On s'intéresse à un évènement A qui peut se décrire en français comme "le résultat est pair". C'est à dire que $A = \{2, 4, 6\}$ (qui est bien à nouveau dans $\mathcal{F}$).

Par additivité de $\mathbb{P}$ on peut calculer

$$\mathbb{P}(A) = \mathbb{P}(\{2\}) + \mathbb{P}(\{4\}) + \mathbb{P}(\{6\}) = 1/6 + 1/6 + 1/6 = 1/2, \tag{1.2.2}$$

la probabilité que le résultat soit pair est $1/2$.

On s'était placé dans une situation très confortable en prenant $\mathcal{F} = \mathcal{P}(\Omega)$ (toutes les situations pouvaient être considérées). On peut imaginer appauvrir la tribu $\mathcal{F}$ et continuer à s'intéresser à l'évènement "le résultat est pair" (seulement pour expliquer les choses, car ici ça n'a pas une grande utilité). Si on pousse les choses à l'extrême la tribu

$$\mathcal{F} = \{\emptyset, \{2, 4, 6\}, \{1, 3, 5\}, \Omega\}$$

convient (le lecteur peut vérifier que c'est une tribu), et c'est la plus petite qui permet de considérer l'évènement $A = \{2, 4, 6\}$.

Par contre on ne peut plus calculer $\mathbb{P}(A)$ par (1.2.2), puisque les singletons $\{i\}$ ne sont pas dans $\mathcal{F}$. Il faudrait que $\mathbb{P}(A)$ nous soit donnée (ou $\mathbb{P}(A^c)$, on ferait alors $\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$). Ou comme le dé est équilibré on peut quand même intuiter que $\mathbb{P}(A) = \mathbb{P}(A^c) = 1/2$.

Noter que si on met sur Ω par exemple la tribu

$$\mathcal{F} = \{\emptyset, \{1\}, \{2, 3, 4, 5, 6\}, \Omega\}$$

alors la question de calculer la probabilité de l'évènement "le résultat est pair", ne fait même plus sens, puisque cet évènement n'est pas dans la tribu.

On peut alors se demander pourquoi on ne prend pas systématiquement $\mathcal{F} = \mathcal{P}(\Omega)$. Cela a l'air en effet plus souple.

Une des raisons est que ce n'est pas toujours possible. Repensons à l'exemple de la mesure de Lebesgue λ sur $([0,1], \mathcal{B}([0,1]))$ rencontré à la fin de la section 1.1. Comme $\lambda([0,1]) = 1$ on peut voir $([0,1], \mathcal{B}([0,1]), \lambda)$ comme un espace de probabilité. Or cette construction n'est possible qu'avec la tribu borélienne $\mathcal{B}([0,1])$, on ne peut pas mettre $\mathcal{P}([0,1])$ à la place (la remarque 1.1.3 reste vraie en remplaçant $\mathbb{R}^d$ par $[0,1]$).

En revanche il est vrai que si Ω est dénombrable il n'y a en général pas de problème de définition d'une mesure de probabilité sur $(\Omega, \mathcal{P}(\Omega))$, on a donc intérêt à prendre $\mathcal{F} = \mathcal{P}(\Omega)$.

Remarque 1.2.3. Plus précisément si Ω est dénombrable et qu'on a une famille sommable (p_ω) telle que $\sum_{\omega \in \Omega} p_\omega = 1$ on peut construire une mesure de probabilité $\mathbb{P}$ sur $(\Omega, \mathcal{P}(\Omega))$ avec $\mathbb{P}(\{\omega\}) = p_\omega$, $\omega \in \Omega$, par la propositon 1.1.2.

On retiendra que dans la très grande majorité des cas on procède comme ceci : on travaille avec une tribu aussi fine (grosse) que l'on peut mais aussi grossière (petite) que l'on doit.

La définition 1.2.1 d'un espace de probabilité, avec sa tribu d'évènements, et sa mesure de probabilités, proposée par Kolmogorov dans les années 1930, donne un cadre axiomatique minimal qui permet au calcul des probabilités de se développer.

Il permet de prendre en compte des situations où on s'intéresse à la répétition infinie d'un certain type d'évènements, comme nous l'illustrons dans l'exemple 1.2.1 ci-après. Pour présenter l'exemple nous avons besoin de définir la notion d'évènements indépendants et au passage nous définissons la probabilité d'un évènement sachant un autre. Jusqu'à la fin du chapitre dans les définitions et résultats on suppose qu'un espace de probabilité $(\Omega, \mathcal{F}, \mathcal{P})$ est donné.

Définition 1.2.2. i) Deux évènements $A, B \in \mathcal{F}$ sont dits indépendants si $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

ii) Soit $(A_i)_{i \in I}$ une famille (finie ou infinie dénombrable) d'évènements de $\mathcal{F}$. Ces évènements sont dits indépendants (on dit parfois mutuellement indépendants) si pour toute partie finie $J \subset I$ on a

$$\mathbb{P}(\bigcap_{i \in J} A_i) = \prod_{i \in J} \mathbb{P}(A_i).$$

Exercice 1.2.1. Montrer que si A et B sont indépendants alors A^c et B^c le sont aussi.

Définition 1.2.3. Soit $B \in \mathcal{F}$ avec $\mathbb{P}(B) > 0$. Pour tout $A \in \mathcal{F}$ on appelle probabilité de A sachant B la quantité

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Exemple 1.2.1. On lance une pièce équilibrée. Puis on recommence, sans jamais s'arrêter. On se propose de montrer que la probabilité qu'on n'aie jamais aucun pile est nulle.

On se place sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ assez riche pour modéliser le problème.

Long aparté : Si on veut décrire en détail l'espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ eh bien on tombe déjà sur une difficulté.

Le Ω le plus petit qu'on puisse prendre c'est $\Omega = \{F, P\}^{\mathbb{N}^*}$. Or cet ensemble n'est pas dénombrable. En effet il est aussi gros que $\{0,1\}^{\mathbb{N}^*}$, qui est lui même aussi gros que $\{0,1\}^{\mathbb{Z}^*}$ qui contient toutes les écritures possibles en binaire d'un nombre réel. En fait $\Omega = \{F, P\}^{\mathbb{N}^*}$ est aussi gros que $\mathbb{R}$. On ne peut donc pas prendre $\mathcal{F} = \mathcal{P}(\Omega)$ et construire $\mathbb{P}$ par la remarque 1.2.3. Il faut à nouveau faire appel à Carathéodory.

Précisons d'abord ce qu'on veut obtenir : notons $F_n = \{\omega \in \Omega : F \text{ apparait au n-ème tirage}\}$, $n \in \mathbb{N}^*$. Supposons pour généraliser la construction qu'on veut que la probabilité d'avoir face soit $p \in [0.1]$. On veut en outre que les F_n soient indépendants (ce seront des évènements une fois l'espace construit). En fait on veut qu'in fine pour I et J ensembles finis d'indices disjoints on ait

$$\mathbb{P}(\bigcap_{i \in I} F_i \bigcap_{j \in J} F_j^c) = p^{|I|}(1-p)^{|J|}. \tag{1.2.3}$$

Effectuons maintenant la construction : on considère donc l'algèbre $\mathcal{A}$ des réunions finies de parties de Ω qui s'écrivent comme $\bigcap_{i\in I} F_i \bigcap_{j\in J} F_j^c$, avec I et J ensembles finis d'indices disjoints. Sur $\mathcal{A}$ on définit l'application $\tilde{\mathbb{P}}$ par

$$\tilde{\mathbb{P}}(\bigcap_{i\in I} F_i \bigcap_{j\in J} F_j^c) = p^{|I|}(1-p)^{|J|}$$

$\mathbb{P}(\emptyset) = 0$ et $\mathbb{P}(\Omega) = 1$ et on demande qu'elle soit additive.

Par le théorème 1.1.1 l'application $\tilde{\mathbb{P}}$ s'étend en une mesure de probabilité $\mathbb{P}$ sur $(\Omega, \mathcal{F} = \sigma(\mathcal{A}))$ qui vérifie par construction (1.2.3).

Il va de soi que dans un exercice de probabilité standard on ne vous demandera de faire une telle construction (!) On partira du principe que l'espace probabilisé existe car on serait capable de le construire.

Sur un tel espace $(\Omega, \mathcal{F}, \mathbb{P})$ (donc!) on considère les évènements F_n, $n \in \mathbb{N}^*$, qui correspondent à "avoir face au n-ème tirage" (ici et après on fait comme si on n'a pas lu le long aparté). Il n'y a pas de raison de supposer que les tirages sont dépendants, on suppose donc que les F_n sont indépendants. De plus comme la pièce est équilibrée on a $\mathbb{P}(F_n) = 1/2$ pour tout $n \in \mathbb{N}^*$.

L'évènement "on n'a jamais aucun pile" c'est $\bigcap_{n\in\mathbb{N}^*} F_n$ (noter qu'il est bien dans $\mathcal{F}$ par stabilité par intersection dénombrable).

L'astuce c'est de remarquer que $\bigcap_{n\in\mathbb{N}^*} F_n = \bigcap_{n\in\mathbb{N}^*} \bigcap_{k=1}^n F_k$ (on laisse le lecteur s'en convaincre).

On remarque alors que la suite $(\bigcap_{k=1}^n F_k)_{n\in\mathbb{N}^*}$ est une suite décroissante d'évènements de $\mathcal{F}$.

Par continuité séquentielle décroissante (proposition 1.1.1-ii)) et indépendance des F_k on a donc

$$\mathbb{P}(\bigcap_{n\in\mathbb{N}^*} F_n) = \mathbb{P}(\bigcap_{n\in\mathbb{N}^*} \bigcap_{k=1}^n F_k) = \lim_{n\to\infty} \mathbb{P}(\bigcap_{k=1}^n F_k) = \lim_{n\to\infty} \prod_{k=1}^n \mathbb{P}(F_k) = \lim_{n\to\infty} \left(\frac{1}{2}\right)^n = 0,$$

CQFD.

Nous introduisons maintenant la notion de propriété vraie *presque sûrement* qui est simplement une traduction de la définition 1.1.4 en vocabulaire probabiliste :

Définition 1.2.4. On dit qu'une propriété $P(\omega)$ est vraie $(\mathbb{P}-)$presque sûrement (p.s.), ou pour $(\mathbb{P}-)$presque tout ω, si l'ensemble $\{\omega \in \Omega : \ P(\omega) \text{ n'est pas vraie}\}$ est ($\mathbb{P}$-)négligeable, i.e. inclus dans $A \in \mathcal{F}$ avec $\mathbb{P}(A) = 0$.

Remarque 1.2.4. On utilisera parfois des variantes équivalentes de la définition 1.2.4.

Par exemple : "On dit qu'une propriété $P(\omega)$ est vraie $(\mathbb{P}-)$presque sûrement (p.s.), ou pour $(\mathbb{P}-)$presque tout ω, si il existe $A \in \mathcal{F}$ avec $\mathbb{P}(A) = 0$ et $P(\omega)$ vraie pour tout $\omega \in A^c$."

Ou encore : "On dit qu'une propriété $P(\omega)$ est vraie $(\mathbb{P}-)$presque sûrement (p.s.), ou pour $(\mathbb{P}-)$presque tout ω, si il existe $\widetilde{\Omega} \in \mathcal{F}$ avec $\mathbb{P}(\widetilde{\Omega}) = 1$ et $P(\omega)$ vraie pour tout $\omega \in \widetilde{\Omega}$."

Ces variantes jettent un éclairage différent sur la définition et sont parfois en pratique plus faciles à vérifier.

Comme pour la définition 1.1.4 s'il n'y a pas d'ambiguité quant à la mesure on dit simplement "presque sûr". Mais il peut arriver qu'on mette diverses mesures de probabilité sur un même espace probabilisable $(\Omega, \mathcal{F})$. Dans ce cas il peut être utile de dire "$\mathbb{P}$-presque sûr", afin d'insister sur la mesure de probabilité utilisée pour vérifier la définition.

Enfin si pour $A \in \mathcal{F}$ on a $\mathbb{P}(A) = 1$ on dit que l'évènement A est presque sûr (noter que dans le définition 1.2.4 $\{\omega \in \Omega : \ P(\omega) \text{ n'est pas vraie}\}$ ou son complémentaire $\{\omega \in \Omega : \ P(\omega) \text{ est vraie}\}$ ne sont pas forcément dans $\mathcal{F}$, d'où l'introduction de la notion d'ensemble négligeable; cependant il arrivera qu'on soit dans la situation plus simple où ils sont directement dans $\mathcal{F}$, ce sont des évènements).

Nous terminons ce chapitre par deux fameux résultats connus sous les noms de lemmes de Borel-Cantelli (mais leur importance justifie qu'on les mette dans un théorème). Pour une suite (A_n) d'évènements nous avons besoin de définir les objets suivants :

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

$$\text{et} \quad \liminf A_n = \bigcup_{n\geq 0}\bigcap_{k\geq n} A_k.$$

Remarque 1.2.5. Remarquons l'analogie avec les définitions de $\limsup$ et $\liminf$ pour une suite réelle (a_n) : par exemple $\limsup a_n = \lim_{n\to\infty}\sup_{k\geq n} a_k = \inf_{n\geq 0}\sup_{k\geq n} a_k$.

Théorème 1.2.1. *Soit (A_n) une suite d'évènements.*
i) **Premier lemme de Borel-Cantelli :** *Si $\sum_{n\geq 0}\mathbb{P}(A_n) < \infty$ alors $\mathbb{P}(\limsup A_n) = 0$.*
ii) **Deuxième lemme de Borel-Cantelli :** *Si les évènements A_n sont indépendants alors on a que $\sum_{n\geq 0}\mathbb{P}(A_n) = \infty$ implique $\mathbb{P}(\limsup A_n) = 1$.*

Démonstration. On a bien sûr que $\Big(\bigcup_{k\geq n} A_k\Big)_n$ est une suite décroissante d'évènements de $\mathcal{F}$. Donc par continuité séquentielle décroissante on a que $\mathbb{P}(\limsup A_n) = \lim_{n\to\infty}\mathbb{P}(\cup_{k\geq n}A_k)$, et il s'agit donc d'identifier la valeur de cette dernière limite dans les points i) et ii).

Pour le point i) : par σ-sous-additivité de $\mathbb{P}$ on a que $\mathbb{P}(\cup_{k\geq n}A_k) \leq \sum_{k\geq n}\mathbb{P}(A_k)$. Or $\sum_{k\geq n}\mathbb{P}(A_k)$ est le reste d'ordre n d'une série convergente par hypothèse. Il tend donc vers zéro quand n tend vers l'infini, ce qui amène le résultat.

Pour le point ii) on remarque d'abord que pour tout $n \in \mathbb{N}$ et tout $N \geq n$ on a

$$\mathbb{P}\Big(\bigcup_{n\leq m\leq N} A_m\Big) = 1 - \mathbb{P}\Big(\bigcap_{n\leq m\leq N} A_m^c\Big) = 1 - \prod_{n\leq m\leq N}\mathbb{P}(A_m^c) = 1 - \prod_{n\leq m\leq N}(1-\mathbb{P}(A_m)),$$

en utilisant l'indépendance des évènements A_m et l'exercice 1.2.1. Puis on utilise le fait qu'on a $1 - x \leq e^{-x}$ pour tout $x \geq 0$, ce qui amène

$$\mathbb{P}\Big(\bigcup_{n\leq m\leq N} A_m\Big) \geq 1 - \prod_{n\leq m\leq N}\exp(-\mathbb{P}(A_m)) = 1 - \exp\Big(-\sum_{n\leq m\leq N}\mathbb{P}(A_m)\Big). \tag{1.2.4}$$

Pour tout $n \in \mathbb{N}$ on a $\lim_{N\to\infty}\sum_{n\leq m\leq N}\mathbb{P}(A_m) = +\infty$ par hypothèse, donc le terme de droite dans (1.2.4) tend vers 1 quand N tend vers l'infini. D'un autre côté $\big(\bigcup_{n\leq m\leq N} A_m\big)_N$ est une suite croissante d'évènements donc par continuité séquentielle croissante le terme de gauche dans (1.2.4) tend vers $\mathbb{P}\Big(\bigcup_{m\geq n} A_m\Big)$ quand N tend vers l'infini.

Finalement pour tout n on a que $\mathbb{P}(\cup_{k\geq n}A_k) = 1$ ce qui implique que $\lim_{n\to\infty}\mathbb{P}(\cup_{k\geq n}A_k) = 1$. □

Commentons ce que nous disent les lemmes de Borel-Cantelli.

Pour commencer il faut bien voir que l'évènement $\limsup A_n$ c'est "une infinité de A_k sont réalisés". En effet soit $\omega \in \limsup A_n$. Pour tout $n \geq 0$ il est dans $\bigcup_{k\geq n} A_k$ c'est à dire qu'on peut trouver au moins un indice $k(\omega) \geq n$ tel que $\omega \in A_{k(\omega)}$. Pour les mêmes raisons on peut trouver un $k'(\omega) > k(\omega)$ tel que $\omega \in A_{k'(\omega)}$ etc...

Donc par exemple le deuxième lemme de Borel-Cantelli nous dit que si les A_n sont indépendants et la série $\sum\mathbb{P}(A_n)$ est divergente alors presque sûrement une infinité de A_k sont réalisés.

Des exemples concrets d'application de ces lemmes seront vus en TD.

Chapitre 2

Variables aléatoires et leurs lois

Dans ce chapitre on introduit la notion de variable aléatoire et de loi de variable aléatoire. Les lois des variables aléatoires peuvent être appréhendées de diverses façons et pour les variables aléatoires dites réelles l'usage de la fonction de répartition est une méthode privilégiée, sur laquelle nous passerons un certain temps. Il sera question de lois à densité, d'espérance et de moments d'une variable aléatoire, donc nous nous reposerons à nouveau beaucoup sur des résultats d'intégration du cours Analyse pour l'ingénieur avancée.

2.1 Variables aléatoires et leurs lois : premières définitions

On rappelle tout d'abord la notion de fonction mesurable.

Définition 2.1.1. Soient $(E_1, \mathcal{E}_1)$ et $(E_2, \mathcal{E}_2)$ deux espaces mesurables. Une fonction $f : E_1 \to E_2$ est dite mesurable si $\forall B \in \mathcal{E}_2$, on a $f^{-1}(B) \in \mathcal{E}_1$.
Si $(E_i, \mathcal{E}_i) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, $i = 1, 2$ on parle de fonction borélienne.

Un espace de probabilités $(\Omega, \mathcal{F}, \mathbb{P})$ est dorénavant donné. Sauf mention contraire les diverses définitions et assertions concernant les variables aléatoires sont données avec cet espace.

Définition 2.1.2. Une variable aléatoire (v.a.) définie sur $(\Omega, \mathcal{F}, \mathbb{P})$, à valeurs dans E (muni d'une tribu $\mathcal{E}$), est une application $X : \Omega \to E$, mesurable de $(\Omega, \mathcal{F})$ vers $(E, \mathcal{E})$. Si $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ on parle de variable aléatoire réelle.

Pour X v.a. à valeurs dans $(E, \mathcal{E})$ et $B \in \mathcal{E}$ on note

$$\{X \in B\} := X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}.$$

On voit l'intérêt de la clause de mesurabilité dans la définition 2.1.2 : l'ensemble $\{X \in B\} \subset \Omega$ est un évènement. On peut donc lui attribuer une probabilité $\mathbb{P}(\{X \in B\})$ que l'on note $\mathbb{P}(X \in B)$, pour alléger.
On va beaucoup dans la suite du cours chercher à évaluer les probabilités d'évènements qui concernent des variables aléatoires.

On note $\sigma(X)$ la plus petite tribu sur Ω qui rend mesurable X.
Traduction : i) quand on dit qu'une tribu $\mathcal{Y}$ rend mesurable X c'est que X est mesurable de $(\Omega, \mathcal{Y})$ vers $(E, \mathcal{E})$.
ii) Quand on dit que $\sigma(X)$ est le plus petite qui réalise cela c'est que pour toute tribu $\mathcal{Y}$ qui rend mesurable X on a $\sigma(X) \subset \mathcal{Y}$ (c'est la plus petite au sens de l'inclusion).
Bien sûr $\mathcal{F}$ rend mesurable X si bien que $\sigma(X)$ est sous-tribu de $\mathcal{F}$. Sur l'existence et l'unicité de $\sigma(X)$ cf Feuille de TD 1 exercice 6. Nous verrons en particulier que $\sigma(X) = \{X^{-1}(B)\}_{B \in \mathcal{E}}$.

Définition 2.1.3. Soit X variable aléatoire définie sur $(\Omega, \mathcal{F}, \mathbb{P})$ à valeurs dans $(E, \mathcal{E})$. On appelle loi de X (sous $\mathbb{P}$) et on note $\mathbb{P}_X$ la mesure image $\mathbb{P} \circ X^{-1}$ définie par

$$\mathbb{P}_X(B) = \mathbb{P} \circ X^{-1}(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}(X \in B), \quad \forall B \in \mathcal{E}. \tag{2.1.1}$$

Remarque 2.1.1. Quand on dit que $\mathbb{P}_X$ est une mesure image cela signifie en particulier que $\mathbb{P}_X$ est une mesure. Or poser (2.1.1) ne permet pas de voir ça très directement, il faut faire une petite vérification : pour toute suite (A_i) d'éléments disjoints de $\mathcal{E}$ on a

$$\mathbb{P}_X\big(\bigcup_i A_i\big) = \mathbb{P}\big(X \in \bigcup_i A_i\big) = \mathbb{P}\big(\bigcup_i \{X \in A_i\}\big) = \sum_i \mathbb{P}(X \in A_i) = \sum_i \mathbb{P}_X(A_i),$$

où à la deuxième égalité on a utilisé que $f^{-1}(\cup_i A_i) = \cup_i f^{-1}(A_i)$ pour toute fonction f (la pré-image de la réunion est la réunion des pré-images), et à la troisième la σ-additivité de $\mathbb{P}$. On a donc montré que $\mathbb{P}_X$ est σ-additive.

Par ailleurs $\mathbb{P}_X(E) = \mathbb{P}(X \in E) = 1$. Donc (cf remarque 1.2.2) on a que $\mathbb{P}_X$ est en fait une mesure **de probabilité** sur $(E, \mathcal{E})$.

Remarque 2.1.2. Dans certains ouvrages vous verrez la mesure image plutôt notée $\mathbb{P} \circ X$. C'est une affaire de convention.

Une variable aléatoire X prend donc certaines valeurs $X(\omega)$ selon l'état du monde ω correspondant. Elle permet en quelque sorte de voir les états du monde à travers cette quantité d'intérêt. Et la loi $\mathbb{P}_X$ de la variable X conduit à s'intéresser au poids $\mathbb{P}_X(B)$ d'ensembles B d'états possibles de X, c'est à dire qu'on considère $\mathbb{P}$ à travers son image (son transport) par la v.a. X.

Par exemple imaginons qu'on lance une pièce équilibrée deux fois de suite et qu'on s'intéresse à la variable aléatoire $X =$"nombre de face obtenu". On peut prendre

$$\Omega = \{FF, FP, PF, PP\} \quad \text{et} \quad \mathcal{F} = \mathcal{P}(\Omega),$$

on a $\mathbb{P}(\omega) = 1/4$ pour tout $\omega \in \Omega$. Ainsi X est à valeur dans $E = \{0, 1, 2\}$ qu'on équipe de $\mathcal{E} = \mathcal{P}(E)$, et on a plus précisément $X(FF) = 2$, $X(FP) = X(PF) = 1$ et $X(PP) = 0$ (on peut vérifier sans difficulté que X est ainsi mesurable, ce qui autorisait bien à l'appeler variable aléatoire). On a par exemple

$$\mathbb{P}_X(1) = \mathbb{P}(X = 1) = \mathbb{P}(FP \cup PF) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

En complétant avec le fait que $\mathbb{P}_X(0) = \mathbb{P}_X(2) = \frac{1}{4}$ on caractérise complètement $\mathbb{P}_X$ en tant que mesure de probabilité sur $(E, \mathcal{E})$.

Exercice 2.1.1. Identifier $\sigma(X)$ dans l'exemple ci-dessus (c'est une tribu strictement plus petite que $\mathcal{P}(\Omega)$, et elle contient 8 sous-parties de Ω).

On rappelle que la mesure de Dirac en $x_0 \in E$ sur $(E, \mathcal{E})$ est définie par

$$\forall B \in \mathcal{E}, \quad \delta_{x_0}(B) = \begin{cases} 1 & \text{si} \quad x_0 \in B \\ 0 & \text{si} \quad x_0 \notin B \end{cases} \tag{2.1.2}$$

(cf Exemple 2.2-1) dans le polycopié de Analyse pour l'ingénieur avancée; notons que la donnée de (2.1.2) ne nous dit pas clairement que δ_{x_0} est une mesure sur $(E, \mathcal{E})$, c'est un exercice que de le vérifier). Cette mesure vérifie en particulier $\delta_{x_0}(E) = 1$ (c'est une mesure de probabilité) et $\delta_{x_0}(\{x_0\}) = 1$ (elle charge le singleton $\{x_0\}$).

Définition 2.1.4. On dit qu'une loi P sur $(E, \mathcal{E})$ est discrète si elle s'écrit $P = \sum_{i \in I} p_i \delta_{x_i}$ avec I ensemble fini ou infini dénombrable d'indices, $x_i \in E$, $\forall i \in I$, $p_i \geq 0$, $\forall i \in I$ et $\sum_{i \in I} p_i = 1$ (de sorte que $P(E) = 1$).

Ainsi on parlera de v.a. X *de loi discrète*, cela signifie simplement que $\mathbb{P}_X$ est une loi discrète.

L'ensemble des x_i dans la définition ci-dessus tels que le p_i correspondant vérifie $p_i > 0$ est appelé *support de la loi* de X (en général on choisit I et les x_i tels que tous les p_i sont strictement positifs, mais sait-on jamais).

Dans cette situation, pour tout x_i dans le support de la loi de X on a simplement

$$\mathbb{P}(X = x_i) = \mathbb{P}_X(\{x_i\}) = p_i > 0 \tag{2.1.3}$$

(la donnée des $\mathbb{P}(X = x_i)$ suffit à décrire complètement la loi de X).

Exemple 2.1.1. 1) La loi de Bernoulli de paramètre $p \in [0, 1]$, notée $\mathrm{Ber}(p)$. Elle est définie par : si $P = \mathrm{Ber}(p)$ on a $P = p\delta_1 + (1-p)\delta_0$ (on peut la voir comme définie sur $\mathbb{R}$ ou $\{0, 1\}$, son support).

Donc si X est de loi de Bernoulli $\mathrm{Ber}(p)$ (on note $X \sim \mathrm{Ber}(p)$, et cela signifie $\mathbb{P}_X = \mathrm{Ber}(p)$) on a

$$\mathbb{P}(X = 1) = \mathbb{P}_X(1) = p \quad \text{et} \quad \mathbb{P}(X = 0) = 1 - p.$$

2) La loi de Poisson de paramètre $\lambda \in \mathbb{R}_+^*$, notée $\mathcal{P}(\lambda)$. Elle est définie par : si $P = \mathcal{P}(\lambda)$ on a $P = \sum_{n\in\mathbb{N}} e^{-\lambda} \frac{\lambda^n}{n!} \delta_n$ (on peut la voir comme définie sur $\mathbb{R}$ ou $\mathbb{N}$, son support).

Si $X \sim \mathcal{P}(\lambda)$ on a

$$\mathbb{P}(X = n) = e^{-\lambda} \frac{\lambda^n}{n!}, \quad \forall n \in \mathbb{N}.$$

Noter qu'on a bien $\sum_n e^{-\lambda} \frac{\lambda^n}{n!} = e^{-\lambda} \sum_n \frac{\lambda^n}{n!} = e^{-\lambda} e^{\lambda} = 1$ ce qui assure qu'on a bien affaire à une loi de probabilité.

La notion de loi discrète s'oppose à la notion de loi à densité que nous verrons dans la section ci-après (pour la définir nous devons au préalable investir dans quelques notions d'intégration). Cependant nous verrons aussi qu'il n'y a pas de raison qu'une loi soit soit discrète, soit à densité, elle peut mêler les deux aspects (cf théorème 2.2.3).

Nous terminons cette section par quelques réflexions sur les relations entre l'espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ où est définie une v.a. X et son espace d'état $(E, \mathcal{E})$.

Déjà remarquons que $(E, \mathcal{E}, \mathbb{P}_X)$ est bien sûr un espace de probabilité (remarque 2.1.1). Si bien que si tout ce qui nous intéresse c'est de définir une variable aléatoire de loi $\mathbb{P}_X$ l'usage de l'espace $(\Omega, \mathcal{F}, \mathbb{P})$ est un peu artificiel (voire "mythique" si on cite [1] p44).

Pour illustrer notre propos prenons l'exemple simple de la loi de Bernoulli (exemple 2.1.1-1)). Supposons qu'on veuille définir sur un espace de probabilité une v.a. X de loi $P = \mathrm{Ber}(p)$ avec $p \in [0, 1]$, l'espace d'état naturel de X sera $(\{0, 1\}, \mathcal{P}(\{0, 1\}))$.

Sur l'espace $(\{0, 1\}, \mathcal{P}(\{0, 1\}))$ on définit la mesure de probabilité P par $P(1) = p$ et $P(0) = 1 - p$ (proposition 1.1.2 et remarque 1.2.3).

On définit ensuite $X : \{0, 1\} \to \{0, 1\}$ par $X(\omega) = \omega, \ \forall \omega \in \{0, 1\}$ (i.e. X est la fonction identité de $\{0, 1\}$ vers lui-même).

Par construction $(\{0, 1\}, \mathcal{P}(\{0, 1\}), P)$ est un espace de probabilité. L'application X est bien sûr mesurable de $(\{0, 1\}, \mathcal{P}(\{0, 1\}))$ vers lui-même, c'est une variable aléatoire définie sur $(\{0, 1\}, \mathcal{P}(\{0, 1\}), P)$ et à valeurs dans $(\{0, 1\}, \mathcal{P}(\{0, 1\}))$. Et on a pour tout $B \in \mathcal{P}(\{0, 1\})$:

$$P_X(B) = P(X \in B) = P(\{\omega \in \{0, 1\} : X(\omega) \in B\}) = P(\{\omega \in \{0, 1\} : \omega \in B\}) = P(B)$$

c'est à dire que la loi de X sous P est bien $P = \mathrm{Ber}(p)$.

Ici l'espace $(\{0, 1\}, \mathcal{P}(\{0, 1\}))$ est donc à la fois l'espace probabilisable sur lequel on définit X et son espace d'état. Une telle construction est dite *canonique*, on ne peut pas faire plus transparent.

Bien sûr la construction canonique n'est pas la seule envisageable et on aurait pu définir X sur une infinité d'autres espaces de probabilité.

Par ailleurs des calculs analogues à ci-dessus montrent que de façon générale toute mesure de probabilité $\mathbb{P}$ sur un espace mesurable $(\Omega, \mathcal{F})$ est la loi d'une v.a. X définie sur $(\Omega, \mathcal{F}, \mathbb{P})$ (il suffit de prendre $X = \mathrm{id}$!).

On prendra donc garde au fait que le langage probabiliste a tendance à confondre mesures de probabilités et lois des variables aléatoires. Il faudra adapter son point de vue au contexte (cf p45 dans [1]).

2.2 Lois de probabilité et intégration : lois à densité, variables aléatoires indépendantes, espérance des variables aléatoires

2.2.1 Préambule : quelques résultats d'intégration utilisés par la suite

On aura parfois besoin du résultat technique suivant.

Propriété 2.2.1. Soit (f_n) une suite de fonctions mesurables de $(E, \mathcal{E})$ espace mesurable vers $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.
Alors les fonctions $\inf_n f_n$, $\sup_n f_n$, $\liminf_n f_n$ et $\limsup_n f_n$ sont mesurables.

Remarque 2.2.1. Ci-dessus par exemple $\inf_n f_n$ désigne la fonction $x \mapsto \inf_n f_n(x)$ et $\liminf_n f_n$ désigne la fonction $x \mapsto \sup_n \inf_{k \geq n} f_k(x)$.

Et c'est un exercice de base sur la notion de fonction mesurable que de vérifier la propriété suivante.

Propriété 2.2.2. Soit g mesurable de $(E_1, \mathcal{E}_1)$ vers $(E_2, \mathcal{E}_2)$ et f mesurable de $(E_2, \mathcal{E}_2)$ vers $(E_3, \mathcal{E}_3)$ (les $(E_i, \mathcal{E}_i)$, $i = 1, 2, 3$ sont des espaces mesurables).
Alors $f \circ g$ est mesurable de $(E_1, \mathcal{E}_1)$ vers $(E_3, \mathcal{E}_3)$.

On rappelle la notation suivante : pour $(E, \mathcal{E})$ espace mesurable et $B \in \mathcal{E}$ on note

$$\mathbf{1}_B(x) = \begin{cases} 1 & \text{si } x \in B \\ 0 & \text{sinon.} \end{cases}$$

Le lecteur pourra vérifier aisément que la fonction $\mathbf{1}_B$ est mesurable de $(E, \mathcal{E})$ vers $(\overline{\mathbb{R}}_+, \mathcal{B}(\overline{\mathbb{R}}_+))$ (ou l'espace plus petit $(\{0,1\}, \mathcal{P}(\{0,1\})$ par exemple; mais notre but est plutôt d'aller vers des fonctions mesurables vers $(\overline{\mathbb{R}}_+, \mathcal{B}(\overline{\mathbb{R}}_+)))$.

On appelle fonction positive *étagée* (on dit aussi parfois fonction *simple*) une fonction de E vers $\overline{\mathbb{R}}_+ = [0, +\infty]$ qui s'écrit

$$f = \sum_{i=1}^{n} b_i \mathbf{1}_{B_i}$$

pour $n \in \mathbb{N}^*$, des valeurs $b_i \in [0, +\infty]$, $1 \leq i \leq n$, et des éléments B_i, $1 \leq i \leq n$, de $\mathcal{E}$. Dans cette écriture on utilise la convention que si $x \notin B$ alors $+\infty \mathbf{1}_B(x) = 0$ (ce petit artifice permettra de considérer des variables aléatoires qui valent l'infini sur certains évènements mais pas sur d'autres).

Exercice 2.2.1. Montrer qu'une fonction étagée définie comme ci-dessus est mesurable de $(E, \mathcal{E})$ vers $(\mathbb{R}+, \mathcal{B}(\overline{\mathbb{R}}_+))$.

Indication : on pourra récrire f sous forme dite *canonique* i.e. $f = \sum_{i=1}^{n} b'_i \mathbf{1}_{B'_i}$ avec les B'_i disjoints et les b'_i distincts.

Il arrivera dorénavant qu'on ne précise plus la tribu sur l'espace d'arrivée. C'est que par défaut on considère que c'est la tribu borélienne. On a le résultat suivant.

Proposition 2.2.1. *Soit $f : E \to [0, +\infty]$ une fonction mesurable positive. Il existe une suite $(f_n)_{n \in \mathbb{N}}$ de fonctions étagées positives qui tend en croissant vers f c'est à dire que*
i) $0 \leq f_n(x) \leq f_{n+1}(x)$ pour tout $x \in E$.
ii) On a $\lim_{n \to \infty} f_n(x) = f(x)$ pour tout $x \in E$ (convergence simple).

Démonstration. Théorème 2.2 de Analyse pour l'ingénieur avancée. □

Soit μ une mesure sur $(E, \mathcal{E})$ et on considère jusqu'à la fin de la sous-section l'espace mesuré $(E, \mathcal{E}, \mu)$.

L'intégrale d'une fonction positive étagée $f = \sum_{i=1}^{n} b_i \mathbf{1}_{B_i}$ contre μ, notée $\int_E f d\mu$ ou $\int_E f(x)\mu(dx)$ ou encore $\int_E f(x)d\mu(x)$, est définie par

$$\sum_{i=1}^{n} b_i \mu(B_i) \in \overline{\mathbb{R}}_+ = [0, +\infty].$$

Puis l'intégrale $\int_E f d\mu$ d'une fonction mesurable positive $f : E \to [0, +\infty]$ est définie par

$$\sup\Big\{\int_E \varphi d\mu, \text{ pour } \varphi \text{ étagée vérifiant } \varphi \le f\Big\} \in \overline{\mathbb{R}}_+.$$

L'intégrale au sens de Lebesgue ainsi définie a certaines bonnes propriétés comme la positivité.

Lemme 2.2.1. *Soient $f, g : E \to [0, \infty]$ mesurables. Si $f \le g$ alors*

$$\int_E f d\mu \le \int_E g d\mu.$$

Démonstration. Cf Proposition II.1.4 dans [1]. □

Et on peut alors montrer le fameux résultat suivant.

Théorème 2.2.1 (Théorème de convergence monotone de Beppo-Levi)**.** *Soit (f_n) une suite croissante de fonctions mesurables positives ($f_n : E \to [0, \infty]$ pour tout n). Alors*

$$\lim_{n\to\infty} \int_E f_n d\mu = \int_E \big[\lim_{n\to\infty} f_n\big]\, d\mu.$$

Démonstration. Cf Théorème II.2.1 dans [1]. □

Remarque 2.2.2. Noter que $\lim_{n\to\infty} f_n = \liminf_n f_n = \limsup_n f_n$ et est donc mesurable d'après la propriété 2.2.1, ce qui donne un sens à l'intégrale à droite du signe égalité dans le théorème.

En combinant Beppo-Levi et la proposition 2.2.1 on peut montrer la linéarité de l'intégrale : pour $a, b \in \mathbb{R}$, on a

$$\int_E (af + bg) d\mu = a \int_E f d\mu + b \int_E g d\mu$$

(mais ce n'est pas le seul moyen, il est possible de montrer la linéarité avant Beppo-Levi cf proposition 2.7.2 dans [8]). De Beppo-Levi on tire aussi les "propriétés magiques" suivantes.

Proposition 2.2.2. *i) Pour $A \in \mathcal{E}$ on a*

$$\int_E (+\infty \mathbf{1}_A)\, d\mu = \begin{cases} 0 & si\ \mu(A) = 0 \\ +\infty & si\ \mu(A) > 0 \end{cases}$$

ii) Pour $f : E \to [0, \infty]$ on a $\int_E f d\mu = 0$ si et seulement si $f(x) = 0$ p.p.
iii) Si $\int_E f d\mu < \infty$ alors $f(x) < \infty$ p.p.

Démonstration. i) On voit $(+\infty)\mathbf{1}_A$ comme limite de la suite croissante $(n\mathbf{1}_A)_{n\in\mathbb{N}}$. Notons qu'on a pour tout n que $\int_E (n\mathbf{1}_A) d\mu = n\mu(A)$. Or par Beppo-Levi il vient

$$\int_E (+\infty \mathbf{1}_A)\, d\mu = \lim_{n\to\infty} \int_E (n\mathbf{1}_A) d\mu = \lim_{n\to\infty} n\mu(A) = \begin{cases} 0 & \text{si } \mu(A) = 0 \\ +\infty & \text{si } \mu(A) > 0 \end{cases}$$

ii) On pose $A = \{x \in E : f(x) \neq 0\}$. Pour la condition suffisante il suffit de remarquer que $f(x) = 0$ p.p. signifie $\mu(A) = 0$, on remarque donc que $\int_E f d\mu = \int_E \mathbf{1}_A f d\mu \le \int_E (+\infty \mathbf{1}_A)\, d\mu$ et on conclut par le point i). Pour la condition nécessaire on renvoie au théorème 2.5 du cours Analyse pour l'ingénieur avancée.

iii) Montrons que $\int_E f d\mu < \infty$ implique $f(x) < \infty$ p.p. Il revient au même de montrer que $\mu(\{x \in E : f(x) = +\infty\}) > 0$ implique $\int_E f d\mu = +\infty$. Notons que par mesurabilité de f l'ensemble $A := \{x \in E : f(x) = +\infty\}$ est dans $\mathcal{E}$. On a $f = f(\mathbf{1}_A + \mathbf{1}_{A^c}) \ge +\infty \mathbf{1}_A$. D'où $\int_E f d\mu \ge \int_E (+\infty \mathbf{1}_A) d\mu = \infty$ en utilisant le i). □

Rappelons les notations $f_+(x) = \max(f(x), 0)$ et $f_-(x) = \max(-f(x), 0)$ pour les parties positive et négative d'une fonction.

L'intégrale d'une fonction mesurable $f : E \to [-\infty, \infty]$ (de signe quelconque) peut être définie si elle est intégrable i.e. $\int_E f_+ d\mu < \infty$ et $\int_E f_- d\mu < \infty$ (ce qui est équivalent à $\int_E |f| d\mu < \infty$; noter la différence par rapport aux deux situations précédentes ou l'intégrale peut être définie même dans le cas non intégrable). Si on est dans ce cas on note $f \in \mathcal{L}^1(E, \mathcal{E}, \mu)$ et l'intégrale est définie par

$$\int_E f d\mu = \int_E f_+ d\mu - \int_E f_- d\mu$$

(cf section 2.4 du cours Analyse pour l'ingénieur avancée pour des détails sur ces questions).

L'intégrale ainsi construite est à nouveau linéaire et positive. Elle possède encore bien d'autres propriétés que nous présenterons au fur et à mesure de nos besoins pour ne pas allonger encore cette sous-section.

Exercice 2.2.2. Soit $x_0 \in \mathbb{R}$ on considère la mesure δ_{x_0}. Montrer que pour toute fonction borélienne f on a $\int_{\mathbb{R}} f d\delta_{x_0} = f(x_0)$.

2.2.2 Lois à densité

Nous devons d'abord définir les notions de mesure absolument continue par rapport à une autre et de densité. Nous regroupons tout dans le théorème que voici.

Théorème 2.2.2 (Théorème de Radon-Nikodym)**.** *Soient μ et ν deux mesures finies sur un espace mesurable $(E, \mathcal{E})$. Il y a équivalence entre*

i) la mesure ν est absolument continue par rapport à μ (chose qu'on note $\nu << \mu$ et qui signifie que pour tout $N \in \mathcal{E}$ t.q. $\mu(N) = 0$ on a $\nu(N) = 0$).

ii) Il existe $f : E \to [0, \infty]$ mesurable et intégrable t.q. $\nu(B) = \int_B f d\mu, \quad \forall B \in \mathcal{E}$. La fonction f s'appelle la densité de ν par rapport à μ et est parfois notée $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$.

Démonstration. Cf Théorème 2.35.3 dans [8]. □

Définition 2.2.1. Si une loi P sur $(E, \mathcal{E})$ est absolument continue par rapport à une mesure μ on dit que P est une loi à densité par rapport à μ (et cette densité est $\frac{\mathrm{d}P}{\mathrm{d}\mu}$).

Si une loi P sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ est à densité f par rapport à la mesure de Lebesgue on dit simplement que P est de densité f.

Par extension si X v.a. définie sur $(\Omega, \mathcal{F}, \mathbb{P})$ à valeurs dans $\mathbb{R}^d$ est telle que sa loi $\mathbb{P}_X$ est de densité f, on dit que X est *de loi à densité* f. Notons qu'une telle densité vérifie $\int_{\mathbb{R}^d} f d\lambda = \mathbb{P}_X(\mathbb{R}^d) = 1$ (une densité intègre à un).

Notons de plus que pour tout $B \in \mathcal{B}(\mathbb{R}^d)$ on a $\mathbb{P}(X \in B) = \int_B f(x)\lambda(dx)$.

On appelle le support de la densité f le support de la loi de X.

Exemple 2.2.1. 1) La loi exponentielle de paramètre $\lambda > 0$, notée $\mathcal{E}xp(\lambda)$. Elle est définie par : si $P = \mathcal{E}xp(\lambda)$ on a $P(dx) = \lambda e^{-\lambda x}\mathbf{1}_{\mathbb{R}_+^*}(x)dx$ (on peut la voir comme définie sur $\mathbb{R}$ ou $\mathbb{R}_+^*$, son support).

Donc si $X \sim \mathcal{E}xp(\lambda)$ on a pour tout $B \in \mathcal{B}(\mathbb{R}_+^*)$,

$$\mathbb{P}(X \in B) = \int_B \lambda e^{-\lambda x}\mathbf{1}_{\mathbb{R}_+^*}(x)dx$$

où on a noté dx au lieu de $\lambda(dx)$ pour alléger (par défaut toutes les intégrales contre la mesure de Lebesgue rencontrées seront notées dorénavant ainsi). On remarque que $\int_0^{+\infty} \lambda e^{-\lambda x}dx = 1$, ce qui assure la bonne définition des choses.

2) La loi normale notée $\mathcal{N}(m, \sigma^2)$: son support est $\mathbb{R}$ tout entier et elle est de densité

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

Reprenons la situation où X est de loi $\mathbb{P}_X$ à densité f, sur $\mathbb{R}$ pour fixer les idées. Pout tout $x_0 \in \mathbb{R}$ on a

$$\mathbb{P}(X = x_0) = \mathbb{P}_X(\{x_0\}) = \int_{\{x_0\}} f(x)dx = 0,$$

où à la troisième égalité on a utilisé la proposition 2.2.2 et le fait que $\lambda(\{x_0\}) = \lambda([x_0, x_0]) = x_0 - x_0 = 0$.

A retenir : si la loi de X est à densité alors elle ne charge aucun point, et ce même si la densité vaut l'infini en certains points.

Les lois à densité par rapport à la mesure de Lebesgue se comportent donc d'une façon très différente des lois discrètes (qui chargent des points).

Mais il n'y a pas de raison qu'une loi soit soit à densité, soit discrète. En fait plus précisément on a la résultat suivant.

Théorème 2.2.3. *Soit P une loi de probabilité sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. On a la décomposition suivante*

$$P = P_{ac} + \sum_i p_i \delta_{x_i} + P_s$$

où P_{ac} est absolument continue par rapport à la mesure de Lebesgue et P_s est une mesure singulière, i.e. ni discrète ni absolument continue (i.e. $P(dx) = f(x)dx + \sum_i p_i \delta_{x_i}(dx) + P_s(dx)$ avec f positive mesurable et intégrable et les $p_i \geq 0$).

Démonstration. Cf [1] théorème II.3.4 et commentaires p47. □

Remarque 2.2.3. Il est difficile de décrire une mesure singulière, qui est telle que $P_s(\{x\}) = 0, \forall x \in \mathbb{R}^d$ et pourtant n'admet pas de densité par rapport à la mesure de Lebesgue. Nous éviterons d'en rencontrer en TD, et dans la plupart des application des probabilités et statistiques en ingénierie on n'en rencontrera pas. Une façon de décrire une telle mesure est par sa fonction de répartition qui est du coup une fonction singulière (cf exemple 2.3.1).

Notons que nous nous focaliserons dans ce cours sur des lois sur $\mathbb{R}^d$ et des densités par rapport à la mesure de Lebesgue. Mais la notion de densité d'une loi par rapport à une mesure (éventuellement elle-même une loi de probabilité) peut intervenir sur des espaces beaucoup plus complexes (cf cours de 3A sur les processus à temps continu).

Avant de passer à la sous-section suivante nous donnons une formule de changement de variable pour les densités de lois de variables aléatoires à valeurs dans $\mathbb{R}^d$. En fait c'est sa démonstration qui va reposer sur la formule de changement de variable dans les intégrales multiples que nous rappelons ici (cf théorème 2.17 du cours Analyse pour l'ingénieur avancée) :

Théorème 2.2.4. *Soit $\phi : \mathcal{U}_1 \subset \mathbb{R}^d \to \mathcal{V}_1 \subset \mathbb{R}^m$ un difféomorphisme (i.e. une bijection de classe C^1 t.q. ϕ^{-1} est aussi de classe C^1) et f est une fonction intégrable sur $\mathcal{U}_1$.*

Alors on a

$$\int_{\mathcal{U}_1} f(u)du = \int_{\mathcal{V}_1} f(\phi^{-1}(v))\,|\det \mathrm{Jac}\,\phi^{-1}|(v)\,dv$$

où on rappelle la notation

$$\mathrm{Jac}\,\varphi = \begin{pmatrix} \partial_{v_1}\varphi_1 & \dots & \partial_{v_m}\varphi_1 \\ \vdots & & \vdots \\ \partial_{v_1}\varphi_d & \dots & \partial_{v_m}\varphi_d \end{pmatrix}$$

avec les φ_i définissant les composantes de φ : $\varphi = (\varphi_1, \dots, \varphi_d)$.

On peut alors montrer le résultat suivant.

Théorème 2.2.5. *Soit U une variable aléatoire dont la loi est de support $\mathcal{U} \subset \mathbb{R}^d$ et de densité f. On pose $V = \phi(U)$ où $\phi : \mathcal{U} \subset \mathbb{R}^d \to \mathcal{V} \subset \mathbb{R}^d$ est un difféomorphisme.*

Alors la loi de V est de de densité $v \mapsto |\det \mathrm{Jac}\,\phi^{-1}|(v)\, f \circ \phi^{-1}(v)$ sur le support $\mathcal{V}$.

Démonstration. Soit un borélien $B \subset \mathcal{V} \subset \mathbb{R}^d$. On a

$$\mathbb{P}(V \in B) = \mathbb{P}(\phi(U) \in V) = \mathbb{P}(\,U \in \phi^{-1}(B)\,)$$

par définition de la préimage $\phi^{-1}(B) \subset \mathcal{U}$. Comme U est de loi à densité f on a

$$\mathbb{P}(V \in B) = \int_{\phi^{-1}(B)} f(u)du.$$

On pose le changement de variable bijectif $v = \phi(u)$, $u = \phi^{-1}(v)$, en considérant la restriction $\phi : \mathcal{U}_1 := \phi^{-1}(B) \to \mathcal{V}_1 := B$. Par la formule de changement de variable du théorème 1 on a

$$\mathbb{P}(V \in B) = \int_B |\det \mathrm{Jac}\, \phi^{-1}|(v)\, f \circ \phi^{-1}(v) dv.$$

Comme ce résultat est vrai pour tout borélien $B \subset \mathcal{V}$ on a bien le résultat annoncé (notons en particulier que $\mathbb{P}(V \in \mathcal{V}) = 1 = \int_{\mathcal{V}} |\det \mathrm{Jac}\, \phi^{-1}|(v)\, f \circ \phi^{-1}(v) dv$). □

2.2.3 Variables aléatoires indépendantes

On commence par la définition que voici.

Définition 2.2.2. Soient $\mathcal{F}_1$ et $\mathcal{F}_2$ deux tribus ou algèbres de $(\Omega, \mathcal{F}, \mathbb{P})$ (c'est à dire deux sous-tribus de $\mathcal{F}$ ou deux algèbres de parties de Ω incluses dans $\mathcal{F}$).

On dit que $\mathcal{F}_1$ et $\mathcal{F}_2$ sont indépendantes si pour tout $A \in \mathcal{F}_1$ et $B \in \mathcal{F}_2$ les évènements A et B sont indépendants.

Notons qu'on pourrait généraliser une telle définition à une famille infinie dénombrable $(\mathcal{F}_i)_{i \in I}$ d'algèbres dans l'esprit de la définition 1.2.2-ii) : on dirait "les $\mathcal{F}_i$, $i \in I$, sont indépendantes si pour tout $J \subset I$ fini on a que pour tous $A_i \in \mathcal{A}_i$, $i \in J$, les événements A_i sont mutuellement indépendants".

Généraliser en ce sens alourdit systématiquement les écritures. Nous nous bornerons à donner des définitions d'indépendance deux à deux (indépendance de deux v.a. etc). Nous laissons au lecteur le soin de formuler les possibles généralisations.

L'intérêt de la définition 2.2.2 réside en partie dans le résultat ci-dessous.

Proposition 2.2.3. *Soient $\mathcal{A}_1$ et $\mathcal{A}_2$ deux algèbres indépendantes sur $(\Omega, \mathcal{F}, \mathbb{P})$ alors $\sigma(\mathcal{A}_1)$ et $\sigma(\mathcal{A}_2)$ sont des tribus indépendantes.*

Démonstration. La preuve repose sur le "théorème des classes monotones" que nous avons choisi de ne pas énoncer dans ce polycopié (cf [1] théorème I.3.3 et proposition IV.1.5). □

On peut donner une première définition de l'indépendance entre deux variables aléatoires.

Définition 2.2.3. Soient X et Y deux variables aléatoires définies sur $(\Omega, \mathcal{F}, \mathbb{P})$. On dit que X et Y sont indépendantes si les tribus $\sigma(X)$ et $\sigma(Y)$ sont indépendantes.

Si par exemple X est à valeurs dans $(E, \mathcal{E})$ on sait que $\sigma(X) = \{X^{-1}(B)\}_{B \in \mathcal{E}}$ (exercice 6 de la fiche de TD1). Il est donc clair qu'on a la définition suivante, équivalente à 2.2.3.

Définition 2.2.4. Soient X et Y deux variables aléatoires définies sur $(\Omega, \mathcal{F}, \mathbb{P})$, à valeurs respectivement dans $(E, \mathcal{E})$ et $(E', \mathcal{E}')$. On dit que X et Y sont indépendantes si on a

$$\forall B \in \mathcal{E},\, B' \in \mathcal{E}', \quad \mathbb{P}(X \in B; Y \in B') = \mathbb{P}(X \in B)\mathbb{P}(Y \in B') \tag{2.2.1}$$

(i.e. les évènements $\{X \in B\}$ et $\{Y \in B'\}$ sont indépendants).

La définition 2.2.3 semble un peu ésotérique et la définition 2.2.4 plus naturelle. Cependant attention : il peut y avoir beaucoup d'éléments dans les tribus $\mathcal{E}$ et $\mathcal{E}'$ et vérifier (2.2.1) peut ne pas être aisé.

Il vaut mieux parfois chercher à vérifier (2.2.1) pour des évènements du type $\{X \in I\}$ (avec I intervalle par exemple) qui sont éléments d'algèbres qui engendrent $\sigma(X)$ et conclure par la proposition 2.2.3.

Exercice 2.2.3. Soient X et Y de lois discrètes, définies sur $(\Omega, \mathcal{F}, \mathbb{P})$ et à valeurs respectivement dans $(E, \mathcal{P}(E))$ et $(E', \mathcal{P}(E'))$ (E, E' sont dénombrables).

On a que X et Y sont indépendantes si et seulement si $\mathbb{P}(X = i;\, Y = j) = \mathbb{P}(X = i)\mathbb{P}(Y = j)$ pour tous $i \in E$, $j \in E'$.

Pour décrire ce qui se passe pour les lois à densité nous avons besoin de décrire un cadre plus général et de donner des éléments sur la notion de mesure produit.

Soient $(E, \mathcal{E}, \mu)$ et $(E', \mathcal{E}', \nu)$ deux espaces mesurés. En général l'ensemble des éléments de $E \times E'$ du type $A \times B$ avec $A \in \mathcal{E}$ et $B \in \mathcal{E}'$ ne constitue pas une tribu : on note $\mathcal{E} \otimes \mathcal{E}'$ la plus petite tribu sur $E \times E'$ qui les contient.

Il est alors possible par extension de montrer qu'il existe une unique mesure $\mu \otimes \nu$ sur $(E \times E', \mathcal{E} \otimes \mathcal{E}')$ qui vérifie

$$\forall A \in \mathcal{E},\, \forall B \in \mathcal{E}', \quad \mu \otimes \nu(A \times B) = \mu(A)\nu(B).$$

Cette mesure $\mu \otimes \nu$ est appelée *mesure produit*. Elle intervient dans le fameux théorème de Fubini que nous rappelons ici.

Théorème 2.2.6 (Théorème de Fubini)**.** *Soit f fonction mesurable de $(E \times E', \mathcal{E} \otimes \mathcal{E}')$ vers $(\overline{\mathbb{R}}, \mathcal{B}(\overline{\mathbb{R}}))$ qui est, soit à valeurs dans $[0, \infty]$, soit intégrable. Alors,*

$$\int_{E \times E'} f\, d(\mu \otimes \nu) = \int_E \Big(\int_{E'} f(x, y)\nu(dy) \Big) \mu(dx) = \int_{E'} \Big(\int_E f(x, y)\mu(dx) \Big) \nu(dy).$$

Démonstration. Cf théorème 10.3 dans [4]. □

Nous avons le résultat suivant, pour la démonstration duquel on utilisera le lemme énoncé juste après.

Proposition 2.2.4. *Soient X et Y deux variables aléatoires, à valeurs respectivement dans $(E, \mathcal{E})$ et $(E', \mathcal{E}')$. Le couple (X, Y) peut être considéré comme une v.a. à valeurs dans $(E \times E', \mathcal{E} \otimes \mathcal{E}')$ et X et Y sont indépendantes si et seulement si la loi $\mathbb{P}_{(X,Y)}$ du couple est la loi produit $\mathbb{P}_X \otimes \mathbb{P}_Y$ sur $(E \times E', \mathcal{E} \otimes \mathcal{E}')$.*

Lemme 2.2.2. *Soit $(E_1, \mathcal{E}_1)$ et $(E_2, \mathcal{E}_2)$ deux espaces mesurables, $\mathcal{C} \subset \mathcal{E}_2$ tel que $\sigma(\mathcal{C}) = \mathcal{E}_2$ et $f : E_1 \to E_2$. Si pour tout $B \in \mathcal{C}$ on a $f^{-1}(B) \in \mathcal{E}_1$ alors f est mesurable de $(E_1, \mathcal{E}_1)$ vers $(E_2, \mathcal{E}_2)$.*

Démonstration. On renvoie à [4] théorème 8.1. □

Preuve de la proposition 2.2.4. Pour tous $A \in \mathcal{E}$, $B \in \mathcal{E}'$ on a par mesurabilité de X et Y

$$(X, Y)^{-1}(A \times B) = X^{-1}(A) \cap Y^{-1}(B) \in \mathcal{F}.$$

En utilisant la définition de la tribu $\mathcal{E} \otimes \mathcal{E}'$ et le lemme 2.2.2 on obtient bien que (X, Y) est une v.a. à valeurs dans $(E \times E', \mathcal{E} \otimes \mathcal{E}')$.

Si X et Y sont indépendantes on a pour tous $A \in \mathcal{E}$, $B \in \mathcal{E}'$,

$$\mathbb{P}_{(X,Y)}(A \times B) = \mathbb{P}(\,(X, Y) \in A \times B\,) = \mathbb{P}(X \in A;\, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B) = \mathbb{P}_X(A)\mathbb{P}_Y(B).$$

Par unicité de la mesure produit on a donc $\mathbb{P}_{(X,Y)} = \mathbb{P}_X \otimes \mathbb{P}_Y$.

Réciproquement si $\mathbb{P}_{(X,Y)} = \mathbb{P}_X \otimes \mathbb{P}_Y$ il est clair par définition de la mesure produit que pour $A \in \mathcal{E}$, $B \in \mathcal{E}'$ on a

$$\mathbb{P}(X \in A;\, Y \in B) = \mathbb{P}(\,(X, Y) \in A \times B\,) = \mathbb{P}_X \otimes \mathbb{P}_Y(A \times B\,) = \mathbb{P}_X(A)\mathbb{P}_Y(B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

d'où l'indépendance de X et Y. □

Proposition 2.2.5. *Soit (X, Y) un couple de variables aléatoires réelles.*

On a que X et Y sont indépendantes de loi à densité respectives f et g si et seulement si la loi de (X, Y) sur $\mathbb{R}^2$ admet la densité $f(x)g(y)$ (i.e. $\mathbb{P}_{(X,Y)}(dxdy) = f(x)g(y)dxdy$) et $\int_{\mathbb{R}} f(x)dx = 1$ (qui entraîne automatiquement $\int_{\mathbb{R}} g(x)dx = 1$).

Démonstration. Notons $\mu_f(dx) = f(x)dx$ et $\mu_g(dy) = g(y)dy$ (à ce stade μ_f et μ_g sont des mesures à densité sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Montrons tout d'abord que

$$\mu_f \otimes \mu_g(dxdy) = f(x)g(y)\, dxdy \tag{2.2.2}$$

(i.e. la mesure produit $\mu_f \otimes \mu_g$ a pour densité $f(x)g(y)$).

Soient $A, B \in \mathcal{B}(\mathbb{R})$ on a

$$\begin{aligned}
\mu_f \otimes \mu_g(A \times B) &= \mu_f(A)\mu_f(B) \\
&= \int_A f(x)dx \int_B g(y)dy \\
&= \int_A \Big(\int_B g(y)dy \Big) f(x)\, dx \\
&= \int_A \Big(\int_B f(x)g(y)dy \Big)\, dx \\
&= \int_{A\times B} f(x)g(y)dxdy
\end{aligned}$$

où on a utilisé le théorème de Fubini à la dernière égalité (et la linéarité de l'intégrale par deux fois). A nouveau par unicité de la mesure produit on a bien montré (2.2.2).

Si X et Y sont indépendantes de lois à densité respectives f et g on a bien sûr que f et g intègrent à un et $\mathbb{P}_X = \mu_f$, $\mathbb{P}_Y = \mu_g$. Si de plus X et Y sont indépendantes on a par la proposition 2.2.4 que $\mathbb{P}_{(X,Y)} = \mathbb{P}_X \otimes \mathbb{P}_Y = \mu_f \otimes \mu_g$ et donc $\mathbb{P}_{(X,Y)}$ admet la densité $f(x)g(y)$ sur $\mathbb{R}^2$ par (2.2.2).

Réciproquement supposons que $\mathbb{P}_{(X,Y)}(dxdy) = f(x)g(y)dxdy$ et que $\int_{\mathbb{R}} f(x)dx = 1$. On a en utilisant (2.2.2) que pour tout $B \in \mathbb{R}$

$$\mathbb{P}(Y \in B) = \mathbb{P}((X,Y) \in \mathbb{R} \times B) = \mu_f \otimes \mu_g(\mathbb{R} \times B) = \mu_f(\mathbb{R})\mu_g(B) = \mu_g(B).$$

En prenant $B = \mathbb{R}$ on voit que $\int_{\mathbb{R}} g(y)dy = 1$ et g apparaît comme la densité de la loi de Y.

Par le même type de calcul on montre que f est la densité de la loi de X. Donc $\mathbb{P}_X = \mu_f$, $\mathbb{P}_Y = \mu_g$ et donc par (2.2.2) on a $\mathbb{P}_X \otimes \mathbb{P}_Y = \mathbb{P}_{(X,Y)}$ ce qui entraîne l'indépendance de X et Y par la proposition 2.2.4. □

2.2.4 Espérance des variables aléatoires

Donnons directement la définition de l'espérance d'une variable aléatoire à valeurs dans $[-\infty, \infty] = \{-\infty\} \cup \mathbb{R} \cup \{+\infty\}$ (essentiellement à valeurs réelles donc mais on autorise les valeurs infinies).

Définition 2.2.5. Soit X v.a. définie sur $(\Omega, \mathcal{F}, \mathbb{P})$ à valeurs dans $[-\infty, \infty]$.

Si X est à valeurs positives (i.e. dans $[0, \infty]$) on définit

$$\mathbb{E}(X) := \int_\Omega X d\mathbb{P} = \int_\Omega X(\omega)\mathbb{P}(d\omega).$$

Si X est à valeurs signées (i.e. dans $[-\infty, \infty]$) on dit qu'elle est intégrable si on a à la fois $\mathbb{E}(X_+) < \infty$ et $\mathbb{E}(X_-) < \infty$ (ce qui équivaut à $\mathbb{E}|X| < \infty$). Dans ce cas on note $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ et on définit

$$\mathbb{E}(X) = \mathbb{E}(X_+) - \mathbb{E}(X_-).$$

Pour fixer les idées imaginons que $X = \mathbf{1}_A$ avec $A \in \mathcal{F}$. Par définition de l'intégrale au sens de Lebesgue on a $\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(A)$.

Imaginons maintenant que $X = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$ avec les a_i dans $[0, \infty]$ et les A_i dans $\mathcal{F}$ (on parle de v.a. étagée). Imaginons pour expliquer les choses qu'on a affaire à l'écriture canonique de cette v.a. étagée (les A_i sont disjoints et les a_i distincts; chaque pré-image de a_i est A_i).

On a dans ce cas $\mathbb{E}(X) = \sum_{i=1}^n a_i \mathbb{P}(A_i)$ c'est à dire que l'espérance de X est la somme de ses valeurs pondérées par les probabilités qu'elle prenne ces valeurs : c'est la valeur "espérée" de X.

Il y a diverses manières de calculer une espérance. Assez souvent on utilise le théorème de transfert que voici.

Théorème 2.2.7. *Soit X définie sur $(\Omega, \mathcal{F}, \mathbb{P})$, à valeurs dans $(E, \mathcal{E})$ espace mesurable. Soit $f : E \to \mathbb{R}$ mesurable.*

Si f est à valeurs positives on a

$$\mathbb{E}[f(X)] = \int_{\mathbb{R}} f(x)\mathbb{P}_X(dx).$$

Si f est à valeurs signées on a $f(X) \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ si et seulement si $f \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$. Dans ce cas l'égalité précédente a lieu.

Remarque 2.2.4. On peut remarquer que grâce à la propriété 2.2.2 on peut effectivement considérer $f(X)$ comme une variable aléatoire et il fait sens de calculer son espérance.

Démonstration. On traite d'abord le cas f à valeurs positives. Si $f = \mathbf{1}_B$ avec $B \in \mathcal{B}(\mathbb{R})$ on a le résultat car

$$\mathbb{E}[\mathbf{1}_B(X)] = \mathbb{E}[\mathbf{1}_{\{X \in B\}}] = \mathbb{P}(X \in B) = \mathbb{P}_X(B) = \int_{\mathbb{R}} \mathbf{1}_B(x)\mathbb{P}_X(dx).$$

Donc par linéarité des intégrales on a le résultat pour f fonction borélienne étagée. Si f est mesurable positive on l'approche par une suite croissante $(f_n)_{n\in\mathbb{N}}$ de fonctions étagées positives, en utilisant la proposition 2.2.1. On a donc

$$\begin{aligned}
\mathbb{E}[f(X)] &= \mathbb{E}[\lim_{n\to\infty} f_n(X)] \\
&= \lim_{n\to\infty} \mathbb{E}[f_n(X)] \\
&= \lim_{n\to\infty} \int_{\mathbb{R}} f_n(x)\mathbb{P}_X(dx) \\
&= \int_{\mathbb{R}} \lim_{n\to\infty} f_n(x)\mathbb{P}_X(dx) \\
&= \int_{\mathbb{R}} f(x)\mathbb{P}_X(dx)
\end{aligned}$$

en utilisant le théorème de Beppo-Levi aux deuxième et quatrième égalités.

En appliquant ce résultat à $|f|$ on montre l'équivalence entre $f(X) \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ et $f \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$. Enfin en appliquant dans le cas intégrable le résultat à f_+ et f_- et en utilisant $f = f_+ - f_-$ on montre la dernière partie du théorème. □

Dans le cas d'une variable aléatoire de loi discrète de support S le théorème de transfert donne la formule bien connue (dans le cas positif ou intégrable)

$$\mathbb{E}[f(X)] = \int_{\mathbb{R}} f(x) \sum_{n\in S} p_n \delta_n(dx) = \sum_{n\in S} f(n)\mathbb{P}(X = n),$$

où on a utilisé (2.1.3) et l'exercice 2.2.2.

Dans le cas d'une variable aléatoire de loi à densité g le théorème de transfert donne

$$\mathbb{E}[f(X)] = \int_{\mathbb{R}} f(x)g(x)dx.$$

Par ailleurs en combinant Fubini, la proposition 2.2.4 et le théorème de transfert on peut montrer le résultat bien connu, énoncé dans le cas intégrable pour fixer les idées.

Proposition 2.2.6. *Soient X et Y deux variables aléatoires indépendantes et intégrables telles que XY est intégrable. On a $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.*

Démonstration. On a

$$\begin{aligned}\mathbb{E}(XY) &= \int_{\mathbb{R}^2} xy\mathbb{P}_{(X,Y)}(dxdy) = \int_{\mathbb{R}^2} xy\mathbb{P}_X\otimes\mathbb{P}_Y(dxdy) = \int_{\mathbb{R}}\Big(\int_{\mathbb{R}} xy\mathbb{P}_X(dx)\Big)\mathbb{P}_Y(dy)\\ &= \int_{\mathbb{R}} y\Big(\int_{\mathbb{R}} x\mathbb{P}(dx)\Big)\mathbb{P}_Y(dy) = \Big(\int_{\mathbb{R}} x\mathbb{P}(dx)\Big)\Big(\int_{\mathbb{R}} y\mathbb{P}_Y(dy)\Big) = \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

Le résultat est démontré. □

2.3 Fonction de répartition d'une variable aléatoire réelle

Pour décrire les lois de variables aléatoires réelles (unidimensionnelles donc) la fonction de répartition (f.d.r.) est un outil privilégié. Nous en rappelons tout d'abord la définition.

Définition 2.3.1. Soit X une variable aléatoire réelle définie sur $(\Omega,\mathcal{F},\mathbb{P})$. On appelle fonction de répartition de X la fonction F_X définie par

$$F_X(x) := \mathbb{P}(X\le x),\quad \forall x\in\mathbb{R}. \tag{2.3.1}$$

On a le résultat suivant, pour la démonstration duquel nous aurons besoin du lemme 2.3.1 ci-après.

Théorème 2.3.1. *La fonction de répartition caractérise complètement la loi d'une variable aléatoire réelle. C'est à dire que pour X et Y deux v.a.r. on a $F_X = F_Y$ si et seulement si $\mathbb{P}_X = \mathbb{P}_Y$.*

Lemme 2.3.1. *On note $\mathcal{B}_0$ l'algèbre de toutes les unions finies disjointes d'intervalles de $\mathbb{R}$ de la forme $]x,y]$ avec $-\infty\le x\le y\le+\infty$ (avec la convention que $]x,\infty]=]x,\infty[$; noter aussi que $]x,y]=\emptyset$ si $x=y$). Alors $\sigma(\mathcal{B}_0)=\mathcal{B}(\mathbb{R})$.*

Démonstration. Après s'être assuré que $\mathcal{B}_0$ est bien une algèbre (soin laissé au lecteur), nous procédons par double inclusion.

Tout intervalle ouvert $]a,b[$ s'écrit $\bigcup_{n=N}^{\infty}]a,b-\frac{1}{n}]$, pour N assez grand pour que $a<b-\frac{1}{N}$. Or chaque $]a,b-\frac{1}{n}]$ est dans $\mathcal{B}_0$ de sorte que $\bigcup_{n=N}^{\infty}]a,b-\frac{1}{n}]$ est dans $\sigma(\mathcal{B}_0)$. Ceci montre que $\sigma(\mathcal{B}_0)$ contient tous les intervalles ouverts de $\mathbb{R}$ or tout ouvert de $\mathbb{R}$ s'écrit comme réunion dénombrable d'intervalles ouverts. Donc la tribu $\sigma(\mathcal{B}_0)$ contient tous les ouverts de $\mathbb{R}$ et on a $\mathcal{B}(\mathbb{R})\subset\sigma(\mathcal{B}_0)$ par définition de $\mathcal{B}(\mathbb{R})$.

Enfin soit $]a,b]\in\mathcal{B}_0$ on a $]a,b]=\bigcap_{n=1}^{\infty}]a,b+\frac{1}{n}[$ qui est dans $\mathcal{B}(\mathbb{R})$ par stabilité par intersection dénombrable. On a donc $\mathcal{B}_0\subset\mathcal{B}(\mathbb{R})$ et par suite $\sigma(\mathcal{B}_0)\subset\mathcal{B}(\mathbb{R})$. □

Preuve du théorème 2.3.1. La condition suffisante est évidente, l'enjeu est de prouver la condition nécessaire.

Soient X et Y deux v.a.r. avec $F_X=F_Y$. Comprenons déjà comment par exemple la donnée de F_X permet d'identifier l'action de $\mathbb{P}_X$ sur les éléments de $\mathcal{B}_0$, l'algèbre évoquée dans le lemme 2.3.1.

Tout d'abord (2.3.1) implique pour tous $x\le y$,

$$\mathbb{P}_X(]x,y]) = \mathbb{P}(x<X\le y) = \mathbb{P}(X\le y)-\mathbb{P}(X\le x) = F_X(y)-F_X(x).$$

Ensuite si $B\in\mathcal{B}_0$ est une union finie disjointe $B=\bigcup_{0\le i\le N}]x_i,y_i]$ on a

$$\mathbb{P}_X(B) = \sum_{i=0}^{n}\big(\,F_X(y_i)-F_X(x_i)\,\big).$$

Par suite $\mathbb{P}_X$ et $\mathbb{P}_Y$ coïncident sur $\mathcal{B}_0$. Par le théorème 1.1.1 et le lemme 2.3.1 on peut conclure qu'elles coïncident sur $\mathcal{B}(\mathbb{R})$ tout entier, i.e. $\mathbb{P}_X=\mathbb{P}_Y$. □

On voit donc que la donnée d'une f.d.r. caractérise complètement la loi de probabilité sur $(\mathbb{R},\mathcal{B}(\mathbb{R}))$ qui lui est associée. Il est donc important d'identifier les propriétés analytiques minimales qui font d'une fonction une f.d.r.

Théorème 2.3.2. *Une fonction de répartition F (associée à une loi P sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ par $F(x) = P(]-\infty, x])$) vérifie*

i) F est croissante (au sens large).

ii) F est continue à droite avec limite à gauche en tout point.

iii) On a $\lim_{x\to-\infty} F(x) = 0$ et $\lim_{x\to+\infty} F(x) = 1$.

Réciproquement si une fonction $F : \mathbb{R} \to [0, 1]$ vérifie les points i)-iii) ci-dessus alors elle est la f.d.r. d'une certaine loi P sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Remarque 2.3.1. Dans certains ouvrages il est précisé que F est à valeurs dans $[0, 1]$ mais c'est simplement imposé par les conditions i) et iii).

Démonstration. Le point i) découle du fait qu'une mesure positive est croissante (propriété 1.1.1).

Pour la continuité à droite dans le point ii) on remarque que $]-\infty, x] = \bigcap_{n=1}^{\infty}]-\infty, x + \frac{1}{n}]$ où la suite d'évènements en jeu dans l'intersection est bien sûr décroissante. On a donc

$$F(x) = P\Big(\bigcap_{n=1}^{\infty}]-\infty, x + \frac{1}{n}]\Big) = \lim_n P(]-\infty, x + \frac{1}{n}]) = \lim_n F(x + \frac{1}{n}) = F(x+).$$

Pour la limite à gauche on procède de façon similaire, mais avec une suite croissante cette fois :

$$\lim_n F(x - \frac{1}{n}) = P\Big(\bigcup_{n=1}^{\infty}]-\infty, x - \frac{1}{n}]\Big) = P(]-\infty, x[)$$

Ceci montre que $F(x-)$ existe et vaut $\mathbb{P}(X < x)$ si P est la loi d'une v.a. X.

Pour le point iii) on utilise encore la continuité séquentielle de la mesure P. On a

$$\lim_{x\to-\infty} F(x) = \lim_{n\to\infty} F(-n) = \lim_{n\to\infty} P(]-\infty, -n]) = P\Big(\bigcap_{n=0}^{\infty}]-\infty, -n]\Big) = P(\emptyset) = 0$$

tandis que

$$\lim_{x\to+\infty} F(x) = \lim_{n\to\infty} F(n) = \lim_{n\to\infty} P(]-\infty, n]) = P\Big(\bigcup_{n=0}^{\infty}]-\infty, n]\Big) = P(\mathbb{R}) = 1.$$

Pour la réciproque on définit $F(-\infty) = 0$ et $F(+\infty) = 1$, en accord avec iii). On cherche ensuite à définir $\tilde{P}$ sur $\mathcal{B}_0$, l'algèbre considérée dans le lemme 2.3.1. Pour $B = \bigcup_{0\le i\le N}]x_i, y_i] \in \mathcal{B}_0$ on définit

$$\tilde{P}(B) = \sum_{i=0}^{n} \big(F(y_i) - F(x_i) \big).$$

On voit déjà que $\tilde{P}(\mathbb{R}) = 1$. Il reste à voir que $\tilde{P}$ est additive sur $\mathcal{B}_0$ pour utiliser le théorème 1.1.1 et le lemme 2.3.1 et conclure que $\tilde{P}$ se prolonge en une mesure de probabilité P sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, dont la f.d.r. sera bien sûr F.

Cette démonstration de l'additivité de $\tilde{P}$ sur $\mathcal{B}_0$ est longue et technique. Nous renvoyons à [4], où elle est menée en détails pages 43 et 44. □

Une f.d.r. n'est donc pas nécessairement continue. Plus précisément on a le résultat suivant :

Proposition 2.3.1. *Une f.d.r. admet au plus un nombre fini dénombrable de points de discontinuité.*

Démonstration. Soit F une f.d.r. on note $D_n = \{x \in \mathbb{R} : \ F(x) - F(x-) \ge \frac{1}{n}\}$ pour tout $n \in \mathbb{N}^*$. Puisque $0 \le F \le 1$ on a forcément $\operatorname{card}(D_n) \le n$. L'ensemble des points de discontinuité de F est $\bigcup_{n=1}^{\infty} D_n$ qui est donc dénombrable. □

Exemple 2.3.1. 1) Pour une v.a. X de loi discrète $\mathbb{P}_X = \sum_{i\in I} p_i \delta_{x_i}$ la f.d.r. est constante par morceaux et saute aux points x_i. On a

$$\forall i \in I, \quad F_X(x_i) - F_X(x_i-) = p_i = \mathbb{P}(X = x_i). \tag{2.3.2}$$

Le mieux pour le voir est de faire un dessin, mais on peut aussi le comprendre par le calcul. En fait nous avons déjà remarqué dans la preuve du théorème 2.3.2 qu'on a $F_X(x-) = \mathbb{P}(X < x)$ pour tout $x \in \mathbb{R}$. Donc $F_X(x) - F_X(x-) = \mathbb{P}(X \leq x) - \mathbb{P}(X < t) = \mathbb{P}(X = x)$. Il est donc clair qu'on a (2.3.2). De plus pour tous x, y avec $x_i \leq x < y < x_{i+1}$ pour un certain $i \in I$ on a $F_X(y) - F_X(x) = 0$ et F_X est donc constante sur le segment $[x, y[$ puisqu'elle est croissante.

Réciproquement si une f.d.r. est constante par morceaux elle est associée à une loi discrète (on laisse le lecteur s'en convaincre).

2) Si la loi de X est à densité f alors $F_X(x) = \int_{-\infty}^x f(z)dz$. Si f est continue (continue par morceaux marche aussi) alors F_X est dérivable et sa dérivée est f, par le théorème fondamental de l'analyse.

Réciproquement on effectuera souvent la démarche suivante : Si F_X est dérivable avec une dérivée au moins continue par morceaux, alors F_X peut s'écrire $F_X(x) = \int_{-\infty}^x F_X'(z)dz$ et cela signifie que la loi de X a une densité f que l'on identifie presque partout à F_X'.

3) La troisième situation est moins naturelle et nous ne la rencontrerons ni en cours, ni en TD. Cependant elle existe et nous la mentionnons pour faire écho à la remarque 2.2.3. Il s'agit du cas où la f.d.r. est

i) continue (on a donc $\mathbb{P}(X = x) = F_X(x) - F_X(x-) = 0$ comme dans le cas à densité)

ii) Ne peut pas s'écrire $\int_{-\infty}^x f(z)dz$ avec quelque fonction Lebesgue intégrable f que ce soit.

C'est à dire que la loi de X est singulière au sens du théorème 2.2.3.

On peut être dans cette situation si F_X est une fonction *singulière* (attention : cette terminologie n'est pas totalement instituée). Une telle fonction est continue, dérivable presque partout, de dérivée nulle mais non constante. L'escalier de Cantor (ou escalier du diable) est un exemple d'une telle fonction, qui croît sur un ensemble négligeable de points (ce qui lui permet d'être non constante).

4) Bien sûr si la loi de X comporte une partie discrète et une partie absolument continue (on écarte en unidimensionnel la partie singulière donc) cela se retrouvera dans la f.d.r. qui pourra croître en étant continue sur certaines portions de $\mathbb{R}$ et sauter (toujours vers le haut) en certains points.

La f.d.r. a diverses utilités. Par exemple c'est un des moyens de décrire la convergence en loi que nous verrons au chapitre 4.

Mais on peut s'en servir à des fins de simulation des variables aléatoires, comme l'illustre la proposition suivante.

On rappelle que la loi uniforme sur $(0, 1)$, notée $\mathcal{U}(0, 1)$ est la loi de densité $\mathbf{1}_{(0,1)}(u)$. Bien sûr la f.d.r. d'une telle loi est donnée par

$$F_U(u) = \begin{cases} 0 & \text{si} \quad u \leq 0 \\ u & \text{si} \quad 0 < u < 1 \\ 1 & \text{si} \quad u \geq 1. \end{cases}$$

Proposition 2.3.2. *Soit X variable aléatoire réelle de f.d.r. F_X strictement croissante. Soit $U \sim \mathcal{U}(0, 1)$. Alors $F_X^{-1}(U)$ a même loi que X.*

Démonstration. On pose $Y = F_X^{-1}(U)$. Calculons la f.d.r. de Y. Pour tout $y \in \mathbb{R}$ on a

$$\mathbb{P}(Y \leq y) = \mathbb{P}(F_X^{-1}(U) \leq y) = \mathbb{P}(\,U \leq F_X(y)\,) = F_X(y).$$

C'est à dire que $F_Y = F_X$ et on conclut par le théorème 2.3.1. □

L'usage de la proposition 2.3.2 pour la simulation est donc limpide. On met en oeuvre l'algorithme :

1) Tirer $U \sim \mathcal{U}(0, 1)$.

2) Renvoyer $F_X^{-1}(U)$

En sortie on récupère une simulation (un tirage) d'une v.a. distribuée comme X.

Remarque 2.3.2. Si F_X n'est pas strictement croissante il faut définir son inverse (dite généralisée) de la façon suivante

$$F_X^{-1}(u) = \inf\{x \in \mathbb{R} : \; F_X(x) \geq u\,\}$$

et le résultat de la proposition est en essence le même.

2.4 Espaces L^p et moments des variables aléatoires

Nous avons déjà introduit la notation $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, qui désigne l'ensemble des v.a.r. intégrables définies sur $(\Omega, \mathcal{F}, \mathbb{P})$.

D'une façon générale pour $1 \leq p < \infty$ on désigne par $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ l'ensemble des variables aléatoires X telles que $\mathbb{E}[\,|X|^p\,] < \infty$.

Pour $p = \infty$ la définition est un peu différente : on désigne par $\mathcal{L}^\infty(\Omega, \mathcal{F}, \mathbb{P})$ l'ensemble des variables aléatoires X pour lesquelles on peut trouver un $0 < c < \infty$ (qui dépend de X) tel que $\{|X| > c\}$ est un évènement $\mathbb{P}$-négligeable.

Comme expliqué dans le cours Analyse pour l'ingénieur avancée (les v.a. y sont remplacées par des fonctions mesurables) on équipe les $\mathcal{L}^p$, $1 \leq p < \infty$ de l'application

$$|| \cdot ||_p : X \mapsto \left(\mathbb{E}[\,|X|^p\,] \right)^{1/p}.$$

L'espace $\mathcal{L}^\infty(\Omega, \mathcal{F}, \mathbb{P})$ est quant à lui équipé de

$$|| \cdot ||_\infty : X \mapsto \inf \left\{ c > 0 : \mathbb{P}(|X| > c) = 0 \right\}$$

et $||X||_\infty$ appelé *supremum essentiel* de X.

Ces applications $|| \cdot ||_p$ sont des semi-normes sur les espaces $\mathcal{L}^p$, pour $1 \leq p \leq \infty$. En effet elles sont absolument homogènes et satisfont l'inégalité triangulaire mais la séparation leur fait défaut.

En effet (prenons $1 \leq p < \infty$ pour fixer les idées) si on a $||X||_p = 0$ cela équivaut à $\mathbb{E}[\,|X|^p\,] = 0$ qui n'entraîne que $|X|^p = 0$ p.s. c'est à dire $X = 0$ p.s. (proposition 2.2.2-ii) ; c'est ce que nous avons vu aussi dans la fiche 2 exercice 2). On n'a donc pas $X = 0$.

Cependant la relation $f = g$ p.s. est une relation d'équivalence (cf définition 2.15 dans le cours Analyse pour l'ingénieur avancée pour plus de détails). Si on considère les espaces $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ quotientés par cette relation d'équivalence on obtient des espaces que l'on note $L^p(\Omega, \mathcal{F}, \mathbb{P})$, avec lesquels on travaille désormais.

Les applications $|| \cdot ||_p$ sont désormais des normes sur les L^p. En effet si $||X||_p = 0$ alors $X = 0$ p.s. (quitte à prendre un représentant de X) c'est à dire que X c'est la classe d'équivalence de zéro.

On a le type de résultat suivant.

Théorème 2.4.1. *Soit $1 \leq p \leq \infty$. On a :*

1) L'inégalité de Hölder : pour $f \in L^p$ et $g \in L^q$ avec q conjugué de p, i.e. $\frac{1}{p} + \frac{1}{q} = 1$, on a $fg \in L^1$ et

$$||fg||_1 \leq ||f||_p \, ||g||_q.$$

2) L'inégalité de Minkowski : pour tous $f, g \in L^p$ on a

$$||f + g||_p \leq ||f||_p + ||g||_p.$$

3) $(\, L^p(\Omega, \mathcal{F}, \mathbb{P}) \,, || \cdot ||_p \,)$ est un espace de Banach.

4) Le dual de $(\, L^p(\Omega, \mathcal{F}, \mathbb{P}) \,, || \cdot ||_p \,)$ est $(\, L^q(\Omega, \mathcal{F}, \mathbb{P}) \,, || \cdot ||_q \,)$ avec q conjugué de p i.e. $\frac{1}{p} + \frac{1}{q} = 1$.

5) Dans le cas $p = 2$ l'espace $L^2(\Omega, \mathcal{F}, \mathbb{P})$ équipé du produit scalaire

$$(X, Y) \mapsto \mathbb{E}(XY)$$

est un Hilbert.

Démonstration. Pour ces résultats techniques nous renvoyons à [1] sections II.6 et V.3 et [4] chapitre 22 □

Le point 5) sera utilisé dans les cours de 2A pour définir l'espérance conditionnelle par projection orthogonale.

On n'utilisera désormais quasiment plus la notation $\mathcal{L}^p$ et aura tendance à confondre un élément d'un espace L^p et son représentant (unique à l'égalité p.s. près). Cet abus de langage (léger au demeurant) est communément pratiqué.

Pour $X \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ on appelle $\mathbb{E}[\,|X|^p\,]$ le moment absolu d'ordre p de X. Et on appelle $\mathbb{E}[\,X^p\,]$ le moment d'ordre p de X. On a par exemple le fameux résultat suivant.

Proposition 2.4.1 (Inégalité de Markov). *Soit $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. On a pour tout $t > 0$*

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

Démonstration. On observe que $\mathbf{1}_{X \geq t} \leq \frac{X}{t} \mathbf{1}_{X \geq t} = \frac{|X|}{t} \mathbf{1}_{X \geq t} \leq \frac{|X|}{t}$. En prenant l'espérance de part et d'autre de l'inégalité on obtient le résultat. □

Pour une v.a.r. X de carré intégrable (i.e. $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, on parle aussi de v.a.r. qui a son moment d'ordre 2) on définit la variance de X par

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}[\,|X - \mathbb{E}(X)|^2\,],$$

c'est l'écart quadratique moyen à la moyenne de X. En utilisant la linéarité de l'espérance il est classique que

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2.$$

On a les conséquences suivantes de l'inégalité de Markov.

Corollaire 2.4.1. *1) Si $X \in L^p$ alors on a $\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}|X|^p}{t^p}$.*
2) **Inégalité de Tchebychev :** *pour $X \in L^2$ on a $\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \frac{\mathbb{V}\mathrm{ar}(X)}{t^2}$.*

Démonstration. Pour le point i) il suffit d'appliquer l'inégalité de Markov à $Y = |X|^p$; pour le point ii) on applique le point i) pour $p = 2$ à $Y = |X - \mathbb{E}(X)|$. □

Nous terminons cette section par l'énoncé de l'inégalité de Jensen, qui entraîne une relation d'inclusion entre les espaces L^p.

Théorème 2.4.2. *Soit X une variable aléatoire réelle intégrable et $\varphi : \mathbb{R} \to \mathbb{R}$ une fonction convexe telle que $\varphi(X)$ est aussi intégrable. Alors*

$$\varphi\big(\mathbb{E}(X)\big) \leq \mathbb{E}[\,\varphi(X)\,].$$

Démonstration. Comme φ est convexe, en tout $t \in \mathbb{R}$ le graphe de φ est au-dessus de sa tangente. Plus précisément il existe β_t tel que $\varphi(x) \geq \varphi(t) + \beta_t(x - t)$ pour tout $x \in \mathbb{R}$ (prendre par exemple β_t la dérivée à droite de φ en t).

Appliquons ce fait avec $t = \mathbb{E}(X)$ et $x = X(\omega)$, pour $\omega \in \Omega$. Il vient

$$\varphi(X(\omega)) \geq \varphi\big(\mathbb{E}(X)\big) + \beta_t(X(\omega) - \mathbb{E}(X))$$

et on a ceci pour tout $\omega \in \Omega$. Comme le passage à l'espérance préserve les inégalités larges on a

$$\mathbb{E}[\,\varphi(X)\,] \geq \varphi\big(\mathbb{E}(X)\big) + \beta_t(\mathbb{E}(X) - \mathbb{E}(X)) = \varphi\big(\mathbb{E}(X)\big)$$

ce qui achève la preuve. □

Corollaire 2.4.2. *Soient $1 \leq p \leq q \leq \infty$. On a $L^q(\Omega, \mathcal{F}, \mathbb{P}) \subset L^p(\Omega, \mathcal{F}, \mathbb{P})$.*

Démonstration. Soit $X \in L^q$ on pose $\varphi(x) = x^{q/p}$, qui est une fonction convexe. Par Jensen on a $\varphi(\mathbb{E}|X|^p) \leq \mathbb{E}|X|^q$ c'est à dire $\mathbb{E}|X|^p \leq \big(\mathbb{E}|X|^q\big)^{p/q} < \infty$. □

Chapitre 3

Autour des vecteurs aléatoires

Par vecteur aléatoire on entend par défaut une variable X à valeurs dans $\mathbb{R}^d$. Comme par projection chaque composante de X est aléatoire on peut voir X comme $X = (X_1, \ldots, X_d)$, c'est à dire le regroupement de ses composantes aléatoires (on note les vecteurs de $\mathbb{R}^d$ en ligne tant qu'on n'a pas à faire de manipulation d'algèbre linéaire dessus ; sinon il faut les penser comme des vecteurs colonne pour souci de cohérence).

Nous avons donc déjà rencontré les vecteurs aléatoires quand nous avons parlé de couple (X, Y) de v.a. dans le chapitre 2. Nous avons vu que si X et Y sont indépendants alors la loi du couple est une loi produit, mais attention dans le cas général ce n'est pas forcément le cas. Par exemple si X et Y sont à valeurs dans $\mathbb{R}$ alors la loi du vecteur sera en toute généralité une certaine loi sur $\mathbb{R}^2$, qu'on peut décrire de diverses façons.

Dans ce chapitre on s'intéresse à diverses questions autour des vecteurs aléatoires. Entre autres on introduit les notions de densité marginale et de densité de la loi conditionnelle. Puis on présente la notion de fonction caractéristique, qui permettra de décrire les lois de vecteurs gaussiens, sur lesquels on passe un certain temps.

3.1 Autour des couples de variables aléatoires

Dans cette section on considèrera des couples de variables aléatoires réelles pour présenter les idées. Cependant tout ce que nous allons décrire se généraliserait à des n-uplets de variables aléatoires à valeurs vectorielles.

Un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ est fixé jusqu'à la fin du chapitre, on ne le reprécisera pas toujours.

3.1.1 Sommes de variables aléatoires indépendantes

Nous avons le résultat classique suivant.

Proposition 3.1.1. *Soient X et Y deux variables aléatoires réelles indépendantes. La loi de $X + Y$ est le produit de convolution des lois $\mathbb{P}_X$ et $\mathbb{P}_Y$ de X et Y, qui est défini par*

$$(\mathbb{P}_X * \mathbb{P}_Y)(B) = \int \Big(\int \mathbf{1}_B(x+y) \mathbb{P}_X(dx) \Big) \mathbb{P}_Y(dy), \quad \forall B \in \mathcal{B}(\mathbb{R}).$$

Démonstration. Par la proposition 2.2.4 la loi du couple (X, Y) est la loi produit $\mathbb{P}_X \otimes \mathbb{P}_Y$. On utilise maintenant le théorème de transfert. Pour tout $B \in \mathcal{B}(\mathbb{R})$ on a

$$\mathbb{P}(X + Y \in B) = \mathbb{E}[\mathbf{1}_B(X+Y)] = \int \mathbf{1}_B(x+y) \mathbb{P}_X \otimes \mathbb{P}_Y(dxdy).$$

Il suffit maintenant d'utiliser le théorème de Fubini pour obtenir le résultat annoncé. $\square$

La proposition débouche sur le encore plus fameux corollaire suivant ("la densité de la somme de v.a.r indépendantes est la convolution de leurs densités").

Corollaire 3.1.1. *Soient X et Y v.a.r. indépendantes de lois à densité respectives f et g. Alors $X+Y$ est de loi à densité*

$$f*g(z)=\int_{\mathbb{R}} f(z-y)g(y)dy=\int_{\mathbb{R}} f(x)g(z-x)dx, \quad z\in\mathbb{R}.$$

Démonstration. Soit $B\in\mathcal{B}(\mathbb{R})$. Par la proposition on a

$$\mathbb{P}_{X+Y}(B)=(\mathbb{P}_X*\mathbb{P}_Y)(B)=\int\Big(\int \mathbf{1}_B(x+y)f(x)dx\Big)g(y)dy$$

Pour y fixé on pose le changement de variable $z=x+y$, $x=z-y$, $dx=dz$ dans l'intégrale en x. Il vient

$$\mathbb{P}_{X+Y}(B)=\int\Big(\int \mathbf{1}_B(z)f(z-y)dz\Big)g(y)dy=\int\Big(\int \mathbf{1}_B(z)f(z-y)g(y)dz\Big)dy.$$

On utilise maintenant Fubini pour obtenir

$$\mathbb{P}_{X+Y}(B)=\int\Big(\int \mathbf{1}_B(z)f(z-y)g(y)dy\Big)dz=\int \mathbf{1}_B(z)\Big(\int f(z-y)g(y)dy\Big)dz$$

Cela montre le résultat avec la première expression de $f*g$. La deuxième expression s'obtient avec des idées analogues. $\square$

3.1.2 Densité marginale, densité de la loi conditionnelle

Soit (X,Y) un couple de v.a.r., de loi $\mathbb{P}_{(X,Y)}(dxdy)$ quelconque. On appelle *lois marginales* les lois des composantes X et Y (on parle par exemple de la marge $\mathbb{P}_X$). On pressent qu'à partir de $\mathbb{P}_{(X,Y)}(dxdy)$ on va pouvoir obtenir les marges en intégrant selon une seule des deux variables.

Les choses sont plus faciles à formaliser dans le cas à densité bivariée c'est à dire quand $\mathbb{P}_{(X,Y)}(dxdy)=f(x,y)dxdy$. Il est en effet évident que pour tout $B\in\mathcal{B}(\mathbb{R})$ on a

$$\mathbb{P}(X\in B)=\mathbb{P}(X\in B;\,Y\in\mathbb{R})=\mathbb{P}_{(X,Y)}(B\times\mathbb{R})=\int_{B\times\mathbb{R}} f(x,y)dxdy=\int_B\Big(\int_{\mathbb{R}} f(x,y)dy\Big)dx,$$

où on a utilisé Fubini à la dernière égalité. Il apparaît donc que $x\mapsto\int_{\mathbb{R}} f(x,y)dy$ est la densité de la loi marginale selon X. De même la marge $\mathbb{P}_Y$ va être donnée par $\mathbb{P}_Y(dy)=\Big(\int_{\mathbb{R}} f(x,y)dx\Big)dy$.

Il est courant de poser les notations $f_X(x)=\int_{\mathbb{R}} f(x,y)dy$ et $f_Y(y)=\int_{\mathbb{R}} f(x,y)dx$ pour ces densités marginales.

De plus ces notations vont permettre de décrire la loi de X conditionnellement à $Y=y$ (cette notion fait sens aussi dans le cas discret mais nous nous limitons au cas à densité bivariée pour la présentation).

Pour tout y tel que $f_Y(y)\neq 0$ on pose

$$f_X^{Y=y}(x):=\frac{f(x,y)}{f_Y(y)}.$$

Compte tenu de la définition de $f_Y(y)$ il est clair que $\int f_X^{Y=y}(x)dx=1$ et $f_X^{Y=y}(x)$ définit donc la densité d'une loi appelée loi de X conditionnellement à $Y=y$.

De plus la connaissance de $f_Y(y)$ et de $f_X^{Y=y}(x)$ permet d'accéder à la loi $f(x,y)dxdy$ du couple (X,Y) comme illustré dans l'exercice suivant.

Exercice 3.1.1. Soient U,V indépendantes et de loi commune $\mathcal{U}(0,1)$. Quelle est la loi de UV ?

3.1.3 Espérance conditionnelle

Avertissement : certains ouvrages de L3 font l'impasse sur l'espérance conditionnelle. En effet cette notion est revue de fond en comble au niveau M1 (2A école d'ingé), où on introduit la notion très puissante d'espérance conditionnellement à une tribu. Cette notion permet de prendre en compte tous les types d'espérance conditionnelle imaginables.

Cependant il nous a semblé opportun, pour un cours de première année d'école d'ingénieur, d'introduire tout de même l'espérance conditionnelle, sur des cas bien particuliers. Nous devons la définir dans ces cas sans avoir recours au langage des espérances conditionnellement à une tribu. Nous faisons confiance aux étudiants pour, une fois arrivés en 2A, prendre du recul et faire le lien avec les objets vus cette année.

Nous allons définir trois types d'espérance conditionnelle.

Espérance conditionnellement à un évènement. Soit $A \in \mathcal{F}$ avec $\mathbb{P}(A) > 0$. Soit X v.a. On définit l'espérance de X sachant l'évènement A comme l'intégrale de X contre $\mathbb{P}(\cdot \mid A)$ (cf exercice 2 fiche 1 de TD pour la définition de cette mesure de probabilité). On a

$$\mathbb{E}_A(X) = \mathbb{E}(X|A) := \int_\Omega X(\omega)\mathbb{P}(d\omega \mid A) = \frac{\mathbb{E}(\mathbf{1}_A X)}{\mathbb{P}(A)} = \frac{\int_A X(\omega)\mathbb{P}(d\omega)}{\mathbb{P}(A)}$$

(les deux premiers termes de l'égalité sont parmi les notations possibles pour l'espérance de X sachant A).

Espérance d'une variable conditionnellement à une autre : cas discret. Soient X et Y deux variables aléatoires discrètes. On note S_Y le support de la loi de Y.

On souhaite définir l'espérance de X sachant Y. Pour tout $y \in S_Y$ on définit

$$\mathbb{E}(X|Y = y) := \mathbb{E}(X \mid \{Y = y\}) = \frac{\mathbb{E}[X\mathbf{1}_{\{Y=y\}}]}{\mathbb{P}(Y = y)}.$$

Ici on a utilisé simplement la définition de l'espérance de X conditionnellement à l'événement $\{Y = y\}$. Cela a un sens puisqu'on a pris y dans le support S_Y. En notant ensuite $\varphi(y) = \mathbb{E}(X|Y = y)$, $y \in S_Y$, on définit la v.a.

$$\mathbb{E}(X|Y) := \varphi(Y).$$

Notons que comme φ est mesurable (c'est un exercice de le vérifier) la v.a. $\mathbb{E}(X|Y)$ est $\sigma(Y)$-mesurable.

Espérance d'une variable aléatoire conditionnellement à une autre : cas à densité bivariée. Soit (X, Y) à valeurs dans $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ et de loi à densité bivariée $f(x, y)$.

On a vu que la loi de la marginale Y est à densité $f_Y(y)$ par rapport à la mesure de Lebesgue et donc $\mathbb{P}(Y = y) = 0$ pour tout $y \in \mathbb{R}$. Donc la définition de l'espérance de X sachant Y donnée plus haut n'a pas de sens. On pose donc, pour tout y t.q. $f_Y(y) > 0$,

$$\mathbb{E}(X|Y = y) := \int_{\mathbb{R}} x\, f_X^{Y=y}(x)\, dx.$$

De façon analogue au cas discret on définit ensuite $\varphi(y) = \mathbb{E}(X|Y = y)$, $y \in \mathbb{R}$, puis

$$\mathbb{E}(X|Y) := \varphi(Y).$$

A nouveau, comme φ est mesurable la v.a. $\mathbb{E}(X|Y)$ est $\sigma(Y)$-mesurable.

On a le résultat suivant.

Proposition 3.1.2 (Formule de l'espérance totale)**.** *Que ce soit dans le cas discret ou dans le cas à densité bivariée décrits plus haut on a*

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

Démonstration. On a en utilisant Fubini pour le cas à densité bivariée.

$$\begin{aligned}
\mathbb{E}\big[\mathbb{E}(X|Y)\big] &= \mathbb{E}\big[\,\varphi(Y)\big] \\
&= \int_{\mathbb{R}} \mathbb{E}(X|Y=y) f_Y(y) dy \\
&= \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} x f_X^{Y=y}(x) dx \Big) f_Y(y) dy \\
&= \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} x\, f_X^{Y=y}(x) f_Y(y)\, dx \Big) dy \\
&= \int_{\mathbb{R}\times\mathbb{R}} x\, f(x,y)\, dx dy \\
&= \mathbb{E}\big[\,X\big].
\end{aligned}$$

Le cas discret est laissé au lecteur. □

Nous verrons en TD des exemples d'utilisation de l'espérance conditionnelle et de la formule de conditionnement.

Exercice 3.1.2. Montrer que la v.a. $\mathbb{E}(X|Y)$ vérifie :
i) Elle est $\sigma(Y)$-mesurable.
ii) Pour tout A dans $\sigma(Y)$ on a $\mathbb{E}\big[\mathbf{1}_A \mathbb{E}(X|Y)\big] = \mathbb{E}\big[\mathbf{1}_A\, X\big]$.

Cet exercice permet par anticipation de faire le lien avec la nouvelle notion d'espérance de X sachant Y qui sera vue au niveau M1...

3.2 Fonctions caractéristiques

Nous avons vu au chapitre 2 que la fonction de répartition caractérise la loi d'une variable aléatoire X (cf théorème 2.3.1).

La notion de fonction de répartition existe pour un vecteur aléatoire $X = (X_1, \ldots, X_d)$. C'est une fonction définie par

$$(x_1, \ldots, x_d) \mapsto \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d).$$

Cependant elle est moins commode à manier qu'en unidimensionnel. En fait nous avons besoin pour la suite d'un nouvel outil pour caractériser la loi des vecteurs aléatoires : la fonction caractéristique. On note $\langle \cdot, \cdot \rangle$ le produit scalaire dans $\mathbb{R}^d$.

Définition 3.2.1. Soit $X = (X_1, \ldots, X_d)^T$ un vecteur aléatoire à valeurs dans $\mathbb{R}^d$. On appelle fonction caractéristique de X la fonction à valeurs complexes

$$\varphi_X(t) = \mathbb{E}[\,e^{i\langle t, X\rangle}\,], \quad t \in \mathbb{R}^d.$$

La fonction caractéristique a diverses propriétés. Nous n'évoquons que celles qui seront utilisées dans le poly. Pour commencer la fonction caractéristique caractérise la loi.

Théorème 3.2.1. *Soient X et Y deux vecteurs aléatoires dans $\mathbb{R}^d$ tels que $\varphi_X = \varphi_Y$. Alors $\mathbb{P}_X = \mathbb{P}_Y$.*

Démonstration. La preuve est difficile et combine des arguments d'analyse et le théorème des classes monotones que nous n'avons pas voulu énoncer dans ce cours. On renvoie le lecteur à [1] théorème III.5.2 ou [4] chapitre 14. □

En revanche les deux résultats suivants sont simples à obtenir.

Théorème 3.2.2. *Soient X et Y deux vecteurs aléatoires indépendants dans $\mathbb{R}^d$. Alors $\varphi_{X+Y} = \varphi_X \varphi_Y$.*

Démonstration. Il suffit de remarquer que

$$\varphi_{X+Y}(t) = \mathbb{E}[\, e^{i\langle t, X+Y\rangle}\,] = \mathbb{E}[\, e^{i\langle t, X\rangle} e^{i\langle t, Y\rangle}\,] = \mathbb{E}[\, e^{i\langle t, X\rangle}\,]\mathbb{E}[\, e^{i\langle t, Y\rangle}\,] = \varphi_X(t)\varphi_Y(t).$$

où l'indépendance (ainsi que l'exercice 1 fiche 3) a été utilisée à la troisième égalité. □

Théorème 3.2.3. *Deux vecteurs aléatoires X et Y sont indépendants si et seulement si*

$$\varphi_{(X,Y)}\big((t_1, t_2)\big) = \varphi_X(t_1)\varphi_Y(t_2). \tag{3.2.1}$$

Démonstration. Si X et Y sont indépendants on a

$$\varphi_{(X,Y)}\big((t_1, t_2)\big) = \mathbb{E}[e^{i\langle(t_1,t_2),(X,Y)\rangle}] = \mathbb{E}[e^{i\langle t_1, X\rangle + i\langle t_2, Y\rangle}] = \mathbb{E}[e^{i\langle t_1, X\rangle}]\mathbb{E}[e^{i\langle t_2, Y\rangle}] = \varphi_X(t_1)\varphi_Y(t_2).$$

Réciproquement si on a (3.2.1) alors on a

$$\int\int e^{i\langle t_1,x\rangle + i\langle t_2,y\rangle}\mathbb{P}_{(X,Y)}(dxdy) = \int e^{i\langle t_1,x\rangle}\mathbb{P}_X(dx)\int e^{i\langle t_2,y\rangle}\mathbb{P}_Y(dy) = \int\int e^{i\langle t_1,x\rangle + i\langle t_2,y\rangle}\mathbb{P}_X\otimes\mathbb{P}_Y(dxdy)$$

(on a utilisé Fubini à la deuxième égalité) c'est à dire que les lois $\mathbb{P}_{(X,Y)}$ et $\mathbb{P}_X \otimes \mathbb{P}_Y$ ont même fonction caractéristique et sont donc égales (par le théorème 3.2.1, quitte à le reformuler un peu en termes de lois). Par la proposition 2.2.4 on a donc l'indépendance de X et Y. □

Il y a un lien entre moments d'un vecteur aléatoire et dérivées de la fonction caractéristique.

Précisons ce qu'on entend par moments d'un vecteur aléatoire. On désigne par $|x| = \sqrt{\langle x, x\rangle}$ la norme euclidienne d'un vecteur $x \in \mathbb{R}^d$. Si $\mathbb{E}(|X|^m) < \infty$ pour $m \in \mathbb{N}^*$ on dit que X *admet des moments jusqu'à l'ordre* m (en effet on a alors $\mathbb{E}|X|^k < \infty$ pour tout $1 \leq k \leq m$, par le corollaire 2.4.2). Dans ce cas, pour tout $k \leq m$, pour tous indices $j_1, \ldots, j_k \in \{1, \ldots, d\}$ la variable aléatoire réelle $X_{j_1} \ldots X_{j_k}$ est intégrable, car majorée en valeur absolue par $|X|^k$. La quantité $\mathbb{E}(X_{j_1} \ldots X_{j_k})$ est parfois appelée $(j_1, \ldots, j_k)$-*moment* de X.

Le lien précédemment évoqué est énoncé dans le théorème que voici.

Théorème 3.2.4. *Soit X un vecteur aléatoire admettant des moments jusqu'à l'ordre m. La fonction caractéristique φ_X de X est m fois continuement différentiable et*

$$\partial^k_{j_1\ldots j_k}\varphi_X(t) = i^k\mathbb{E}\big(X_{j_1}\ldots X_{j_k}e^{i\langle t, X\rangle}\big), \quad \forall 1 \leq k \leq m, \ \ \forall j_1, \ldots, j_k \in \{1, \ldots, d\}.$$

En particulier $\partial^k_{j_1\ldots j_k}\varphi_X(0) = i^k\mathbb{E}\big(X_{j_1}\ldots X_{j_k}\big)$ pour tout $1 \leq k \leq m$.

Démonstration. Cf [4] théorème 13.2. □

Un exemple bien connu de fonction caractéristique est celui de la loi normale $\mathcal{N}(0, 1)$: si X suit une telle loi on a

$$\varphi_X(t) = e^{-t^2/2}.$$

Exercice 3.2.1. Montrer que si $X \sim \mathcal{N}(m, \sigma^2)$ on a $\varphi_X(t) = e^{imt - t^2\sigma^2/2}$.

3.3 Vecteurs gaussiens

3.3.1 Préambule : notion de covariance

On rappelle que pour $X, Y \in L^2$ on note

$$\mathbb{C}\mathrm{ov}(X, Y) = \mathbb{E}\Big[\,(X - \mathbb{E}(X))\,(Y - \mathbb{E}(Y))\,\Big]$$

la covariance de X avec Y. Remarquons que cette quantité existe car $X - \mathbb{E}(X)$ et $Y - \mathbb{E}(Y)$ sont dans L^2 et qu'on a à notre disposition l'inégalité de Hölder (théorème 2.4.1-1)). Par ailleurs en développant le produit et en utilisant la linéarité de l'espérance on obtient l'expression alternative de la covariance :

$$\mathbb{C}\mathrm{ov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Par ailleurs on a bien sûr $\mathbb{C}\mathrm{ov}(X, X) = \mathbb{V}\mathrm{ar}(X)$.

Exercice 3.3.1. Montrer que $(X, Y) \mapsto \mathbb{C}\mathrm{ov}(X, Y)$ est une application bilinéaire symétrique.

On a le résultat bien connu suivant.

Proposition 3.3.1. *Si X et Y sont indépendantes alors $\mathbb{C}\mathrm{ov}(X, Y) = 0$.*

Cependant attention !

Exercice 3.3.2. Soient X et Y indépendants et de loi commune $\mathcal{B}er(\frac{1}{2})$. Montrer que

$$\mathbb{C}\mathrm{ov}(X + Y, |X - Y|) = 0$$

mais que $X + Y$ et $|X - Y|$ ne sont pas indépendants.

La notion de covariance donne naissance à celle de matrice de variance-covariance d'un vecteur aléatoire.

Définition 3.3.1. Soit $X = (X_1, \dots, X_d)$ un vecteur aléatoire ayant tous ses moments jusqu'à l'ordre 2. On appelle espérance de X et on note $\mathbb{E}(X)$ le vecteur $(\mathbb{E}(X_1), \dots, \mathbb{E}(X_d))^T$, et matrice de variance-covariance (ou matrice de covariance) la matrice $C \in \mathbb{R}^{d \times d}$ définie par

$$C_{ij} = \mathbb{C}\mathrm{ov}(X_i, X_j), \quad 1 \leq i, j \leq d.$$

Nous regroupons dans la proposition suivante quelques résultats élémentaires dont nous aurons besoin.

Proposition 3.3.2. *Soit $X = (X_1, \dots, X_d)$ un vecteur aléatoire ayant tous ses moments jusqu'à l'ordre 2 et C sa matrice de covariance.*

i) la matrice C est symétrique et semi-définie positive i.e. $\xi^T C \xi \geq 0$, pour tout $\xi \in \mathbb{R}^d$.

ii) Soit $A \in \mathbb{R}^{m \times d}$, posons $Y = AX$. Alors Y a aussi tous ses moments d'ordre 2, son espérance est $A\mathbb{E}(X)$ et sa matrice de covariance est donnée par ACA^T.

Démonstration. i) Le caractère symétrique de la matrice C est évident par symétrie de $\mathbb{C}\mathrm{ov}(\cdot, \cdot)$. Pour le caractère semi-défini positif il suffit de remarquer que pour tout $\xi \in \mathbb{R}^d$ on a

$$\mathbb{V}\mathrm{ar}(\xi^T X) = \mathbb{C}\mathrm{ov}\big(\sum_{i=1}^{d} \xi_i X_i \, , \, \sum_{j=1}^{d} \xi_j X_j \big) = \sum_{i,j=1}^{d} \xi_i \xi_j \mathbb{C}\mathrm{ov}(X_i, X_j) = \xi^T C \xi$$

grâce à la bilinéarité de $\mathbb{C}\mathrm{ov}(\cdot, \cdot)$ (exercice 3.3.1). Or une variance est toujours positive, ce qui amène le résultat.

ii) En utilisant la notion de norme de matrice on a $|Y| \leq |A| \, |X|$ et donc $\mathbb{E}|Y|^2 \leq |A|^2 \, \mathbb{E}|X|^2 < \infty$.

Ensuite soient $1 \leq i, j \leq m$ on a

$$\begin{aligned}
\mathbb{C}\mathrm{ov}(Y_i, Y_j) &= \mathbb{C}\mathrm{ov}\big(\textstyle\sum_{k=1}^{d} A_{ik} X_k \, , \, \sum_{l=1}^{d} A_{jl} X_l \big) \\
&= \sum_{k,l=1}^{d} A_{ik} \mathbb{C}\mathrm{ov}(X_k, X_l) A_{jl} \\
&= \sum_{k=1}^{d} A_{ik} \sum_{l=1}^{d} C_{kl} A^T_{lj} \\
&= \sum_{k=1}^{d} A_{ik} (CA^T)_{kj} \\
&= (ACA^T)_{ij},
\end{aligned}$$

ce qui amène le résultat voulu. $\square$

3.3.2 Vecteurs gaussiens

Définition 3.3.2. Un vecteur aléatoire $X = (X_1, \dots, X_d)^T$ à valeurs dans $\mathbb{R}^d$ est dit gaussien si pour tout $a = (a_1, \dots, a_d)^T \in \mathbb{R}^d$ la variable aléatoire réelle $\langle a, X\rangle = a^T X = \sum_{i=1}^d a_i X_i$ est une variable aléatoire réelle gaussienne (i.e. de loi $\mathcal{N}(m, \sigma^2)$ avec $m \in \mathbb{R}$ et $\sigma^2 \geq 0$, le cas $\sigma^2 = 0$ étant dit dégénéré).

Exercice 3.3.3. Montrer que toute transformation linéaire d'un vecteur gaussien est encore un vecteur gaussien.

Nous avons le résultat central suivant qui permet d'appréhender les lois de vecteurs gaussiens et de leurs transformations affines.

Théorème 3.3.1. *Un vecteur aléatoire $X = (X_1, \dots, X_d)^T$ est gaussien si et seulement si sa fonction caractéristique est de la forme*

$$\varphi_X(t) = \exp\big(i\langle t, \mu\rangle - \frac{1}{2} t^T Q t\big), \quad t \in \mathbb{R}^d, \tag{3.3.1}$$

où $\mu \in \mathbb{R}^d$ et $Q \in \mathbb{R}^{d\times d}$ est une matrice symétrique semi-définie positive.

Dans ce cas le vecteur μ est la moyenne de X (i.e. $\mu_i = \mathbb{E}(X_i)$ pour tout i) et Q est sa matrice de covariance.

Démonstration. Pour la condition suffisante on suppose qu'on a (3.3.1) et pour $a = (a_1, \dots, a_d)^T \in \mathbb{R}^d$ quelconque on cherche à montrer que $Y = \langle a, X\rangle$ est de loi normale (unidimensionnelle). On cherche à identifier la loi de Y en calculant sa fonction caractéristique. On a pour tout $v \in \mathbb{R}$

$$\varphi_Y(v) = \mathbb{E}[e^{iv\langle a,X\rangle}] = \mathbb{E}[e^{i\langle va,X\rangle}] = \varphi_X(va) = \exp\big(iv\langle a, \mu\rangle - \frac{v^2}{2} a^T Q a\big)$$

Or $\langle a, \mu\rangle$ est une certaine valeur réelle et $a^T Q a \geq 0$ donc $a^T Q a$ peut être considérée comme une variance. On reconnait donc là la fonction caractéristique d'une loi gaussienne 1D (cf exercice 3.2.1).

Pour calculer la moyenne et la covariance de X on utilise le théorème 3.2.4. Les dérivées d'ordre un et deux ont respectivement la forme

$$\partial_j \varphi_X(t) = \varphi_X(t)\big(i\mu_j - \sum_{m=1}^d t_m Q_{mj}\big)$$

et

$$\partial^2_{kj} \varphi_X(t) = \varphi_X(t)\Big[\big(i\mu_k - \sum_{m=1}^d t_m Q_{mk}\big)\big(i\mu_j - \sum_{m=1}^d t_m Q_{mj}\big) - Q_{kj}\Big],$$

si on les prend en $t = 0$ cela donne $\partial_j \varphi_X(0) = i\mu_j$ et $\partial^2_{kj}\varphi_X(0) = -\mu_k\mu_j - Q_{kj}$. Cela amène bien $\mu_k = \mathbb{E}(X_k)$ pour tout i d'une part et $\mu_j\mu_k + Q_{kj} = \mathbb{E}(X_j X_k)$ pour tous k, j d'autre part, c'est à dire $\mathbb{C}\mathrm{ov}(X_j, X_k) = Q_{kj}$.

Pour la condition nécessaire pour $a \in \mathbb{R}^d$ quelconque on considère $Y = \langle a, X\rangle$ dont la loi est gaussienne, en particulier uniquement déterminée par sa moyenne et sa variance que nous calculons. On note μ la moyenne de X et Q sa matrice de covariance. La moyenne de Y est donnée par

$$\mathbb{E}(Y) = \mathbb{E}(\sum_{i=1}^d a_i X_i) = \sum_{i=1}^d a_i \mu_i = \langle a, \mu\rangle.$$

Et on calcule la variance de $Y = a^T X$ en invoquant la proposition 3.3.2-ii), il vient :

$$\mathbb{V}\mathrm{ar}(Y) = a^T Q a = \langle a, Qa\rangle.$$

Finalement, $\varphi_Y(v) = e^{iv\langle a,\mu\rangle - \frac{v^2}{2}\langle a,Qa\rangle}$, $v \in \mathbb{R}$ et en particulier $\varphi_Y(1) = e^{i\langle a,\mu\rangle - \frac{1}{2}\langle a,Qa\rangle}$. Mais $\varphi_Y(1) = \mathbb{E}[e^{i\langle a,X\rangle}] = \varphi_X(a)$, d'où le résultat. □

La loi sur $\mathbb{R}^d$ d'une vecteur gaussien est donc entièrement déterminée par la donnée de son vecteur moyenne μ et de sa matrice de covariance Q (puisque sa fonction caractéristique n'est définie que par ces deux quantités). On note celle loi $\mathcal{N}(\mu, Q)$.

Une fois le théorème 3.3.1 établi toutes les propriétés des vecteurs gaussiens viennent assez naturellement. Nous en évoquons certaines d'entre elles.

La propriété suivante dit que pour obtenir un vecteur gaussien il suffit de stocker des v.a.r gaussiennes indépendantes dans un vecteur.

Propriété 3.3.1. Soient $X_1, \ldots, X_d$ des v.a.r. indépendantes de lois respectives $\mathcal{N}(\mu_i, \sigma_i^2)$, $1 \le i \le d$. Alors le vecteur $X = (X_1, \ldots, X_d)^T$ est gaussien, de moyenne $\mu = (\mu_1, \ldots, \mu_d)^T$ et de matrice de covariance

$$Q = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_d^2 \end{pmatrix}. \tag{3.3.2}$$

Démonstration. Comme les X_i sont indépendants, on a en utilisant le théorème 3.2.3 que

$$\begin{aligned} \varphi_X(t_1, \ldots, t_n) &= \textstyle\prod_{j=1}^d \varphi_{X_j}(t_j) \\ &= \textstyle\prod_{j=1}^d e^{i\mu_j t_j - t_j^2\sigma_j^2/2} \\ &= \exp\big(i \textstyle\sum_{j=1}^d \mu_j t_j - \frac{1}{2}\sum_{j=1}^d t_j^2 \sigma_j^2\big) \\ &= e^{i\langle t, \mu\rangle - \frac{1}{2}\langle t, Qt\rangle} \end{aligned}$$

avec la matrice diagonale Q définie dans l'énoncé de la propriété. On conclut par le théorème 3.3.1. $\square$

Proposition 3.3.3. *Soit $X = (X_1, \ldots, X_d)^T$ un vecteur gaussien. Ses composantes X_j sont indépendantes si sa matrice de covariance Q est diagonale.*

Démonstration. Supposons que Q a la forme (3.3.2). Alors à partir de (3.3.1) on voit que $\varphi_X(t) = \prod_{j=1}^d \varphi_{X_j}(t_j)$ avec $\varphi_{X_j}(t_j) = e^{i\mu_j t_j - t_j^2\sigma_j^2/2}$ pour tout j. On conclut à l'indépendance des X_j par le théorème 3.2.3. $\square$

Remarquons la différence avec le cas général qui a été discuté dans l'exercice 3.3.2 : dans le cas général "covariance nulle" n'implique pas "indépendance". Mais dans le cas des composantes d'un vecteur gaussien si !

Considérons maintenant un vecteur gaussien $X = (X_1, \ldots, X_d)^T$ de loi $\mathcal{N}(0, I_d)$ (on a noté I_d la matrice identité de taille d). Les composantes X_j étant indépendantes et de densité $\frac{1}{\sqrt{2\pi}}\exp(-\frac{x_j^2}{2})$ il est clair par la proposition 2.2.5 que la densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$ de la loi de X est donnée par

$$\frac{1}{(2\pi)^{d/2}} \exp\Big(-\frac{|x|^2}{2}\Big), \quad x = (x_1, \ldots, x_d) \in \mathbb{R}^d.$$

La densité de la loi de tout vecteur gaussien va alors s'obtenir comme corollaire du théorème suivant. On rappelle que toute matrice symétrique semi-définie positive Q peut s'écrire $Q = AA^T$ (plus précisément $Q = PDP^T$ avec D diagonale avec v.p. positives et P orthogonale, avec $PP^T = I_d$; donc poser $A = P\sqrt{D}$ convient). On considère jusqu'à la fin du chapitre que Q est de plein rang.

Théorème 3.3.2. *Soit X vecteur gaussien de taille d et de loi $\mathcal{N}(\mu, Q)$. On considérant Y de loi $\mathcal{N}(0, I_d)$ et en formant $\mu + AY$ où $Q = AA^T$ on obtient un vecteur gaussien qui a même loi que X.*

Démonstration. On traite d'abord le cas $\mu = 0$. On a que AY est un vecteur gaussien (exercice 3.3.3) de moyenne le vecteur nul. Pour identifier sa variance il suffit d'appliquer la proposition 3.3.2-ii) : cette variance est $AA^T = Q$. Donc $\mu + AY$ a pour fonction caractéristique

$$\mathbb{E}[e^{i\langle t, \mu + AY\rangle}] = e^{i\langle t, \mu\rangle}\mathbb{E}[e^{i\langle t, AY\rangle}] = e^{i\langle t, \mu\rangle} e^{-\frac{1}{2}\langle t, Qt\rangle} = \exp\big(i\langle t, \mu\rangle - \frac{1}{2} t^T Q t\big)$$

et on a bien que la loi de $\mu + AY$ est une $\mathcal{N}(\mu, Q)$, celle de X. $\square$

Corollaire 3.3.1. *Soit X vecteur gaussien de taille d et de loi $\mathcal{N}(\mu, Q)$. La densité de la loi de X est donnée par*

$$\frac{1}{(2\pi)^{d/2}\sqrt{\det Q}} \exp\Big(-\frac{1}{2}\langle (x-\mu), Q^{-1}(x-\mu)\rangle\Big), \quad x = (x_1, \dots, x_d) \in \mathbb{R}^d. \tag{3.3.3}$$

Démonstration. On considère à nouveau l'écriture $Q = AA^T$ et on pose $\phi(y) = \mu + Ay$, $y \in \mathbb{R}^d$. On a $\phi^{-1}(x) = A^{-1}(x-\mu)$ et $\mathrm{Jac}(\phi^{-1}) = A^{-1}$. En utilisant le théorème 2 de la fiche de TD 2 on a donc que la densité de $\phi(Y)$ où $Y \sim \mathcal{N}(0, I_d)$ est donnée par

$$\begin{aligned}
&\frac{1}{(2\pi)^{d/2}} \exp\Big(-\frac{1}{2}\langle A^{-1}(x-\mu), A^{-1}(x-\mu)\rangle\Big)|\det(A^{-1})| \\
&= \frac{1}{(2\pi)^{d/2}|\det(A)|} \exp\Big(-\frac{1}{2}\langle (x-\mu), (A^{-1})^T A^{-1}(x-\mu)\rangle\Big) \\
&= \frac{1}{(2\pi)^{d/2}\sqrt{\det(Q)}} \exp\Big(-\frac{1}{2}\langle (x-\mu), Q^{-1}(x-\mu)\rangle\Big).
\end{aligned}$$

A la dernière égalité on a utilisé le fait que $(A^{-1})^T A^{-1} = (A^T)^{-1}A^{-1} = (AA^T)^{-1} = Q^{-1}$ et $\det(Q) = \det(AA^T) = \det(A)\det(A^T) = |\det(A)|^2$. Pour conclure il suffit d'utiliser le théorème pour remarquer que $\phi(Y)$ a la loi de X. $\square$

Moralité : Pour simuler un vecteur de taille d de loi $\mathcal{N}(\mu, Q)$ il suffit de :
i) Simuler d v.a.r. gaussiennes $\mathcal{N}(0, 1)$ indépendantes et les stocker dans un vecteur Y.
ii) Former $X = \mu + AY$ avec $Q = AA^T$ (en pratique on fabrique souvent A par la décomposition de Choleski).

On obtient ainsi un vecteur dont la densité de probabilité est donnée par (3.3.3).

Remarque 3.3.1. Si Q n'est pas de plein rang la loi du vecteur gaussien $\mathcal{N}(\mu, Q)$ peut toujours être considérée, moyennant quelques aménagements. Cf [4] pages 135-136.

Chapitre 4

Convergence des variables aléatoires

Dans ce chapitre on présente les différents modes de convergence des (suites de) variables aléatoires : convergence presque sûre, convergence en probabilité, convergence L^p et convergence en loi. Les liens entre ces divers modes de convergence sont expliqués. A la fin du chapitre on aborde la loi forte des grands nombres et le théorème limite-centrale, qui mettent en jeu deux de ces modes de convergence.

4.1 Convergence presque sûre, en probabilité et pour la norme L^p

Pour introduire la notion de convergence presque sûre revenons à notre exemple 1.2.1, où on tire à pile ou face indéfiniment. Assez naturellement on voudrait que

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{F_i}(\omega) = \frac{1}{2} \tag{4.1.1}$$

pour tout $\omega \in \Omega$ (la fréquence empirique de face tend vers $1/2$ quand le nombre de tirages dans la somme tend vers l'infini, et ce quel que soit l'état du monde).

Or c'est impossible : l'état du monde "que des piles jusqu'à l'infini" aboutit à ce que la limite dans (4.1.1) vaut zéro. Plus généralement les états du monde où il y a un nombre finis de face aboutiront au même résultat : zéro.

Cependant ces états du monde défavorables, même s'ils existent dans Ω, sont contenus dans un évènement négligeable : ce qu'on verra quand on aura montré la loi forte des grands nombres c'est que la convergence (4.1.1) a lieu pour presque tout ω. En d'autres termes la suite $\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{F_i}\right)_{n\geq 1}$ converge vers $1/2$ presque sûrement.

Nous allons maintenant formaliser cette définition. Dans tout le chapitre un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ est fixé et sauf mention contraire c'est sur cet espace que sont définies les variables aléatoires considérées. Quand on parle de suite de variables aléatoires $(X_n)_{n\in\mathbb{N}}$ cela signifie dans cette section que, pour tout $n \in \mathbb{N}$, X_n est une variable aléatoire définie sur $(\Omega, \mathcal{F}, \mathbb{P})$ à valeurs dans $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ (on se limite aux v.a.r. pour fixer les idées).

Définition 4.1.1. On dit que (X_n) suite de v.a.r. définies sur $(\Omega, \mathcal{F}, \mathbb{P})$ converge p.s. vers X v.a.r. définie sur le même espace, et on note

$$X_n \xrightarrow[n\to\infty]{\text{p.s.}} X,$$

si et seulement si il existe $A \in \mathcal{F}$ t.q. $\mathbb{P}(A) = 0$ et

$$\forall \omega \in A^c, \quad X_n(\omega) \xrightarrow[n\to\infty]{} X(\omega).$$

Remarque 4.1.1. Notons qu'en fait l'ensemble $\{\omega \in \Omega : \lim_n X_n(\omega) = X(\omega)\}$ est bien dans $\mathcal{F}$ (ainsi que son complémentaire) car on peut remarquer qu'il est égal à

$$C = \bigcap_{\ell \in \mathbb{N}^*} \bigcup_{n \in \mathbb{N}} \bigcap_{k \geq n} \{ |X_k - X| \leq \frac{1}{\ell} \}$$

(qui est effectivement dans $\mathcal{F}$ on utilisant la mesurabilité des X_n et de X et des arguments de stabilité par réunion et intersection dénombrable). En effet si ω est dans C alors pour tout $\ell \in \mathbb{N}$ il existe $N(\omega) \in \mathbb{N}$ tel que $\forall k \geq N(\omega)$ on a $|X_k(\omega) - X(\omega)| \leq 1/\ell$ ce qui signifie que $\lim_{n\to\infty} X_n(\omega) = X(\omega)$. Cela montre $C \subset \{\omega \in \Omega : \lim_n X_n(\omega) = X(\omega)\}$ et il est facile de voir l'inclusion inverse.

C'est pourquoi certains ouvrages (par exemple [1]) posent simplement comme définition de la convergence presque sûre : $\mathbb{P}(\{\lim_n X_n = X\}) = 1$. Ce qui joue le rôle de A dans la définition 4.1.1 c'est alors tout simplement $\{\lim_n X_n = X\}^c$. Dans la suite nous jonglerons un peu avec ces variantes de la définition.

Identifions tout de suite des outils qui vont nous permettre plus loin de montrer des convergences presque sûres.

Lemme 4.1.1. *Soient (X_n) une suite de v.a.r et X une v.a.r. On a que $X_n \xrightarrow[n\to\infty]{\text{p.s.}} X$ si et seulement si*

$$\forall \varepsilon > 0, \quad \mathbb{P}\big(\bigcap_{n \in \mathbb{N}} \bigcup_{k \geq n} \{ |X_k - X| > \varepsilon \}\big) = 0.$$

Démonstration. Pour la condition nécessaire : par hypothèse il existe A avec $\mathbb{P}(A) = 0$ et $\lim_n X_n(\omega) = X(\omega)$ pour tout $\omega \in A^c$.

Soit $\varepsilon > 0$. Soit $\omega \in \bigcap_{n\in\mathbb{N}} \bigcup_{k\geq n} \{ |X_k - X| > \varepsilon \}$. On a que pour tout $n \in \mathbb{N}$, il existe $k(\omega) \geq n$ tel que $|X_{k(\omega)}(\omega) - X(\omega)| > \varepsilon$. C'est donc que $X_n(\omega)$ ne converge pas vers $X(\omega)$ et donc $\omega \in A$.

On a donc montré l'inclusion $\bigcap_{n\in\mathbb{N}} \bigcup_{k\geq n} \{ |X_k - X| > \varepsilon \} \subset A$ qui entraîne

$$\mathbb{P}\big(\bigcap_{n \in \mathbb{N}} \bigcup_{k \geq n} \{ |X_k - X| > \varepsilon \}\big) = 0.$$

Pour la condition suffisante on a en utilisant l'hypothèse et la σ-sous-additivité de $\mathbb{P}$ que

$$\mathbb{P}\Big(\bigcup_{\ell \in \mathbb{N}^*} \bigcap_{n \in \mathbb{N}} \bigcup_{k \geq n} \{ |X_k - X| > \frac{1}{\ell} \}\Big) \leq \sum_{\ell \in \mathbb{N}^*} \mathbb{P}\Big(\bigcap_{n \in \mathbb{N}} \bigcup_{k \geq n} \{ |X_k - X| > \frac{1}{\ell} \}\Big) = 0.$$

Donc $\mathbb{P}\Big(\bigcap_{\ell\in\mathbb{N}^*} \bigcup_{n\in\mathbb{N}} \bigcap_{k\geq n} \{ |X_k - X| \leq \frac{1}{\ell} \}\Big) = 1$, ce qui signifie que $X_n \xrightarrow[n\to\infty]{\text{p.s.}} X$ d'après la remarque 4.1.1. $\square$

Lemme 4.1.2. *Une condition suffisante pour avoir $X_n \xrightarrow[n\to\infty]{\text{p.s.}} X$ est*

$$\forall \varepsilon > 0, \quad \sum_{n \geq 0} \mathbb{P}(|X_n - X| > \varepsilon) < \infty.$$

Démonstration. Par le premier lemme de Borel-Cantelli on a

$$\forall \varepsilon > 0, \quad \mathbb{P}\big(\bigcap_{n \in \mathbb{N}} \bigcup_{k \geq n} \{ |X_k - X| > \varepsilon \}\big) = 0.$$

et on conclut par le lemme 4.1.1. $\square$

Le convergence presque sûre est encore trop forte pour certaines situations. Nous définissons deux autres modes de convergence.

Définition 4.1.2. Soient (X_n) suite de v.a.r et X v.a.r définies $(\Omega, \mathcal{F}, \mathbb{P})$. On dit que (X_n) converge en probabilités vers X et on note

$$X_n \xrightarrow[n\to\infty]{\mathbb{P}} X,$$

si

$$\forall \varepsilon > 0, \quad \mathbb{P}(|X_n - X| > \varepsilon) \xrightarrow[n\to\infty]{} 0.$$

Définition 4.1.3. Soient (X_n) suite de v.a.r et X v.a.r définies $(\Omega, \mathcal{F}, \mathbb{P})$ et toutes dans $L^p(\Omega, \mathcal{F}, \mathbb{P})$. On dit que (X_n) converge en moyenne L^p vers X et on note

$$X_n \xrightarrow[n\to\infty]{L^p} X$$

si $||X_n - X||_p \to 0$, quand $n \to \infty$.

Remarque 4.1.2. La convergence en moyenne L^2 est aussi appelée convergence en moyenne quadratique.

Il existe des liens entre les trois modes de convergence que nous venons de définir, la convergence en probabilité s'avérant être le plus faible. Nous présentons certains de ces liens sans prétendre à l'exhaustivité.

Tout d'abord nous avons :

Proposition 4.1.1. *Si $X_n \xrightarrow[n\to\infty]{\text{p.s.}} X$ alors $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$.*

Démonstration. Supposons que $X_n \xrightarrow[n\to\infty]{\text{p.s.}} X$. Par le lemme 4.1.1 et par continuité séquentielle décroissante de $\mathbb{P}$ nous avons

$$\forall \varepsilon > 0, \quad \lim_{n\to\infty} \mathbb{P}\big(\bigcup_{k\geq n} \{ \, |X_k - X| > \varepsilon \, \} \big) = 0. \tag{4.1.2}$$

En remarquant que pour tout n et tout $\varepsilon > 0$ on a

$$\mathbb{P}\big(\sup_{k\geq n} |X_k - X| > \varepsilon \big) \leq \mathbb{P}\big(\bigcup_{k\geq n} \{ \, |X_k - X| > \varepsilon \, \} \big)$$

l'équation (4.1.2) implique que

$$\forall \varepsilon > 0, \quad \lim_{n\to\infty} \mathbb{P}\big(\sup_{k\geq n} |X_k - X| > \varepsilon \big) = 0.$$

Or pour tout $\varepsilon > 0$ on a bien sûr $\mathbb{P}(|X_n - X| > \varepsilon) \leq \mathbb{P}\big(\sup_{k\geq n} |X_k - X| > \varepsilon \big)$ pour tout n, ce qui amène le résultat. □

Ensuite l'inégalité de Markov entraîne naturellement le résultat suivant.

Proposition 4.1.2. *Soit $1 \leq p < \infty$. Soient (X_n) suite de v.a.r. toutes dans L^p et X dans L^p.*

Si $X_n \xrightarrow[n\to\infty]{L^p} X$ alors $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$.

Démonstration. Pour tout $\varepsilon > 0$ on a par le corollaire 2.4.1 que $\mathbb{P}(|X_n - X| > \varepsilon) \leq \frac{\mathbb{E}|X_n - X|^p}{\varepsilon^p}$ pour tout n et le résultat en découle. □

Les réciproques des deux résultats précédents sont fausses en général. De plus la convergence p.s. n'entraîne pas la convergence L^p. Quelques contre-exemples classiques sont mentionnés ci-après.

Exercice 4.1.1. Sur l'espace de probabilité $([0,1], \mathcal{B}([0,1], \lambda)$ on définit la suite de variables aléatoires (X_i) par $X_i(\omega) = \mathbf{1}_{](k-1)/2^n, k/2^n]}(\omega)$ où $n = n(i) = \min\{m : \; i+1 \leq 2^{m+1}\}$ et $k = k(i) = i + 1 - 2^n$.

Montrer que (X_i) tend en probabilité vers zéro mais ne converge pas p.s.

Exemple 4.1.1. Soit $p > 1$. Pour tout n soit X_n une v.a. de loi $(1 - n^{-p})\delta_0 + n^{-p}\delta_n$ i.e. telle que

$$\mathbb{P}(X_n = n) = n^{-p} = 1 - \mathbb{P}(X_n = 0).$$

Pour tout $\varepsilon > 0$ on a $\mathbb{P}(|X_n| > \varepsilon) = n^{-p}$ pour tout $n > \varepsilon$. Donc par le lemme 4.1.2 on a $X_n \xrightarrow[n\to\infty]{\text{p.s.}} 0$.

En revanche on a $\mathbb{E}(|X_n|^p) = \frac{|n|^p}{|n|^p} = 1$ donc (X_n) ne converge pas dans L^p vers zéro.

On a cependant le genre de "réciproque partielle" suivante.

Théorème 4.1.1. *Supposons que* $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$.

Alors il existe une sous-suite (n_k) *telle que* $X_{n_k} \xrightarrow[k\to\infty]{\text{p.s.}} X$.

Démonstration. Grâce à l'hypothèse de convergence en probabilité pour tout k on peut considérer n_k le plus petit entier tel que

$$\mathbb{P}(|X_{n_k} - X| > \frac{1}{k}) \leq \frac{1}{2^k}.$$

Alors $\sum_{k=1}^{\infty} \mathbb{P}(|X_{n_k} - X| > \frac{1}{k})) < \infty$ et par le premier lemme de Borel-Cantelli

$$\mathbb{P}(\limsup\{|X_{n_k} - X| > \frac{1}{k}\}) = 0$$

c'est à dire que l'évènement $\bigcup_k \bigcap_{m\geq k}\{|X_{n_m} - X| \leq \frac{1}{m}\}$ est presque-sûr. Or cet évènement c'est "il existe un rang K tel qu'on a $|X_{n_m} - X| \leq \frac{1}{m}$ pour tout $m \geq K$" qui est inclus dans "X_{n_m} converge vers X" (car être sur cet évènement assure que pour tout $\delta > 0$ on pourra trouver un rang à partir duquel $|X_{n_m} - X| \leq \delta$). Ceci achève la preuve. □

Finalement donnons un exemple bien connu de situation où on a une convergence en probabilité qui a quasiment un sens physique : la fameuse loi faible des grands nombres déjà rencontrée en PS1.

Exemple 4.1.2. Soit (X_i) une suite de v.a.r. indépendantes et identiquement distribuées (i.i.d.), toutes dans L^2. Alors

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}(X_1).$$

En effet on remarque que $\mathbb{E}(\overline{X}_n) = \mathbb{E}(X_1)$ et que $\mathbb{V}\text{ar}(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\text{ar}(X_i) = \frac{\mathbb{V}\text{ar}(X_1)}{n}$ (pour le dernier calcul on a utilisé le fait que la variance de la somme de v.a. indépendantes est la somme de leurs variances, c'est en fait un corollaire de l'exercice 3.3.1 et de la proposition 3.3.1).

Par l'inégalité de Tchebychev (corollaire 2.4.1-2)) on a donc immédiatement

$$\forall \varepsilon > 0, \quad \mathbb{P}\big(|\overline{X}_n - \mathbb{E}(X_1)|\big) \leq \frac{\mathbb{V}\text{ar}(X_1)}{n\varepsilon^2} \xrightarrow[n\to\infty]{} 0.$$

La loi faible des grands nombres nous dit bien que la moyenne empirique de v.a. i.i.d. a une certaine tendance à tendre vers la moyenne mathématique (l'espérance), ce qui est intuitif. Cependant elle n'énonce pas la convergence presque sûre que nous évoquions en introduction de cette section. La démonstration de la loi forte des grands nombres est reportée à la section 4.3.

4.2 Convergence en loi

4.2.1 Quelques compléments

Il sera bien pratique d'avoir à notre disposition le condition nécessaire et suffisante de convergence en probabilité présentée dans le théorème 4.2.1 ci-après.

Théorème 4.2.1. *On a* $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$, *si et seulement si*

$$\lim_{n\to\infty} \mathbb{E}\Big(\frac{|X_n - X|}{1 + |X_n - X|}\Big) = 0.$$

Démonstration. Sans perte de généralité on suppose X centrée et on cherche à montrer que $X_n \xrightarrow[n\to\infty]{\mathbb{P}} 0$, équivaut à $\lim_{n\to\infty} \mathbb{E}\Big(\frac{|X_n|}{1+|X_n|}\Big) = 0$.

On suppose d'abord que $X_n \xrightarrow[n\to\infty]{\mathbb{P}} 0$. On a pour tout $\varepsilon > 0$

$$\frac{|X_n|}{1+|X_n|} \leq \frac{|X_n|}{1+|X_n|}\mathbf{1}_{|X_n|>\varepsilon} + \varepsilon\mathbf{1}_{|X_n|\leq\varepsilon} \leq \mathbf{1}_{|X_n|>\varepsilon} + \varepsilon,$$

donc

$$\mathbb{E}\Big(\frac{|X_n|}{1+|X_n|}\Big) \leq \mathbb{E}(\mathbf{1}_{|X_n|>\varepsilon}) + \varepsilon = \mathbb{P}(|X_n| > \varepsilon) + \varepsilon.$$

En prenant les limites il vient

$$\lim_{n\to\infty} \mathbb{E}\Big(\frac{|X_n|}{1+|X_n|}\Big) = \limsup_{n\to\infty} \mathbb{E}\Big(\frac{|X_n|}{1+|X_n|}\Big) \leq \varepsilon.$$

Comme ε est arbitrairement petit il vient $\lim_{n\to\infty} \mathbb{E}\Big(\frac{|X_n|}{1+|X_n|}\Big) = 0$.

Réciproquement supposons que $\lim_{n\to\infty} \mathbb{E}\Big(\frac{|X_n|}{1+|X_n|}\Big) = 0$ et fixons $\varepsilon > 0$. La fonction $f(x) = \frac{x}{1+x}$ est strictement croissante donc

$$\frac{\varepsilon}{1+\varepsilon}\mathbf{1}_{|X_n|>\varepsilon} \leq \frac{|X_n|}{1+|X_n|}\mathbf{1}_{|X_n|>\varepsilon} \leq \frac{|X_n|}{1+|X_n|}.$$

En prenant les espérances, puis les limites, on arrive à

$$\frac{\varepsilon}{1+\varepsilon}\lim_{n\to\infty} \mathbb{P}(|X_n| > \varepsilon) \leq \lim_{n\to\infty} \mathbb{E}\Big(\frac{|X_n|}{1+|X_n|}\Big) = 0.$$

Comme ε est foxé on conclut que $\lim_{n\to\infty} \mathbb{P}(|X_n| > \varepsilon) = 0$. On a montré que $X_n \xrightarrow[n\to\infty]{\mathbb{P}} 0$. □

Nous utiliserons aussi le fameux théorème de convergence dominée de Lebesgue (théorème 2.9 du poly de Analyse pour l'ingénieur avancée), que nous reformulons ici avec un langage probabiliste.

Théorème 4.2.2. *Soit (X_n) une suite de v.a.r. qui admet une domination, i.e. il existe Y v.a.r. positive et intégrable telle que $|X_n| \leq Y$ p.s. pour tout $n \in \mathbb{N}$.*
Si $X_n \xrightarrow[n\to\infty]{\text{p.s.}} X$ alors X est intégrable et $\lim_n \mathbb{E}(X_n) = \mathbb{E}(X)$.

4.2.2 Convergence en loi : définition et quelques caractérisations équivalentes

Pour fixer les idées imaginons que nous avons une suite de v.a.r. (X_n) et X une v.a.r. définies sur $(\Omega, \mathcal{F}, \mathbb{P})$. Dire que (X_n) converge en loi vers X cela va être dire que la suite des lois $\mathbb{P}_{X_n}$ converge vers $\mathbb{P}_X$ en un sens que nous allons préciser.

Comme cette convergence ne met en jeu que les lois notons d'emblée qu'on pourrait être dans la situation où les v.a. X_n ne sont pas toutes définies sur le même espace. A l'extrême on pourrait imaginer que X est définie sur $(\Omega, \mathcal{F}, \mathbb{P})$ et que chaque v.a. X_n est définie sur un espace $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ différent. Cela n'empêcherait pas d'avoir la convergence en loi de (X_n) vers X en ce sens que les lois $\mathbb{P}^n_{X_n}$ convergeraient vers $\mathbb{P}_X$.

Dans ce qui suit nous n'allons pas nous acharner à alourdir les notations en ce sens mais c'est quelque chose qu'il faut avoir à l'esprit, et nous laisserons autant que possible de l'espace dans les définitions pour cette situation.

Définition 4.2.1. Soient (P_n) une suite de lois de probabilité sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ et P une loi de probabilité sur le même espace. On dit que (P_n) converge étroitement vers P et on note $P_n \rightharpoonup P$, si pour toute fonction f continue bornée sur $\mathbb{R}$ on a $\int f\, dP_n \to \int f\, dP$ quand $n \to \infty$.

La définition est ainsi car l'ensemble des probabilités sur $\mathbb{R}$ est dans le dual de l'ensemble des fonctions continues bornées sur $\mathbb{R}$. La convergence étroite est donc une sorte de convergence $*$-faible. Divers outils d'analyse peuvent être donc utilisés pour étudier les convergences étroites. Nous ne rentrerons pas dans les détails dans ce cours.

Définition 4.2.2. Soit (X_n) une suite de variables aléatoires à valeurs dans $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ et X une v.a. à valeurs dans le même espace. Notons (P_n) la suite des lois des X_n et P celle de X (ce sont toutes des mesures de probabilité $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$). On dit que (X_n) converge en loi vers X et on note $X_n \xrightarrow[n\to\infty]{\mathcal{L}} X$, si $P_n \rightharpoonup P$.

Nous avons vu respectivement dans les chapitres 2 et 3 que la f.d.r. et la fonction caractéristique caractérisent la loi d'une v.a.r. Il n'est donc pas surprenant qu'elles aident aussi à caractériser la convergence en loi. Les choses sont formalisées dans le théorème suivant.

Théorème 4.2.3. *On a équivalence entre :*
i) $X_n \xrightarrow[n\to\infty]{\mathcal{L}} X$
ii) $\varphi_{X_n}(t) \xrightarrow[n\to\infty]{} \varphi_X(t)$ *pour tout* $t \in \mathbb{R}$.
iii) $F_{X_n}(x) \xrightarrow[n\to\infty]{} F_X(x)$ *en tout point* x *de continuité de* F_X.

Démonstration. Pour la preuve nous renvoyons à [1] théorème V.4.1. □

Le point ii) va jouer un rôle crucial dans la preuve du théorème de la limite centrale.

Dans le cas où toutes les variables aléatoires sont définies sur le même espace de probabilité on a la caractérisation suivante, qui sert fréquemment.

Proposition 4.2.1. *Soient* (X_n) *une suite de v.a.r. et* X *une v.a.r., définies sur* $(\Omega, \mathcal{F}, \mathbb{P})$. *Alors* $X_n \xrightarrow[n\to\infty]{\mathcal{L}} X$ *si et seulement si pour toute fonction continue bornée* $f : \mathbb{R} \to \mathbb{R}$ *on a* $\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$ *quand* $n \to \infty$.

Démonstration. Par définition la convergence en loi c'est la convergence étroite $P_n \rightharpoonup P$. Or pour toute fonction f continue bornée (donc mesurable et intégrable) on a $\int f(x)\,\mathbb{P}_{X_n}(dx) \to \int f(x)\,d\mathbb{P}_X(dx)$ si et seulement si $\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$, par le théorème de transfert (théorème 2.2.7). □

4.2.3 lien avec les autres modes de convergence

Bien sûr on ne peut comparer la convergence en loi avec les modes de convergence de la section 4.1 que si toutes les v.a. en jeu sont définies sur le même espace. Dans ce cas on se doute que la convergence en loi est la plus faible de toutes.

Encore une fois la question est vaste et on ne prétend pas à l'exhaustivité. On a par exemple le résultat suivant :

Proposition 4.2.2. *La convergence presque sûre entraîne la convergence en loi.*

Démonstration. Supposons que $X_n \xrightarrow[n\to\infty]{\text{p.s.}} X$.

Soit f une fonction continue bornée. Il est assez facile de vérifier que $f(X_n) \xrightarrow[n\to\infty]{\text{p.s.}} f(X)$ (cf [4] théorème 17.5-(a) pour les détails). Or f est bornée, c'est à dire $|f(x)| \leq M$ pour tout x, avec un certain $0 \leq M < \infty$. Bien sûr la constante M est intégrable (elle est dans $L^1(\Omega, \mathcal{F}, \mathbb{P})$). Par convergence dominée (théorème 4.2.2) on a donc que $\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$, et on conclut par la proposition 4.2.1. □

Une autre stratégie pour montrer la proposition 4.2.2 aurait été de combiner la proposition 4.1.1 et le résultat suivant.

Théorème 4.2.4. *Si on a* $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ *alors* $X_n \xrightarrow[n\to\infty]{\mathcal{L}} X$.

Démonstration. Cf [1] pages 122-123. □

On a une réciproque (très !) partielle du résultat précédent.

Théorème 4.2.5. *Soient* (X_n) *une suite de v.a.r. et* X *une v.a.r., définies sur* $(\Omega, \mathcal{F}, \mathbb{P})$. *Si* $X_n \xrightarrow[n\to\infty]{\mathcal{L}} X$ *et si* X *est p.s. égale à une constante* c *alors* $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X = c$.

Démonstration. Supposons que $X = c$ p.s. La fonction $f(x) = \frac{|x-c|}{1+|x-c|}$ est bornée et continue. On a donc par la proposition 4.2.1 que $\mathbb{E}f(X_n) \to \mathbb{E}f(X) = \mathbb{E}f(c) = 0$. On a donc

$$\lim_{n\to\infty} \mathbb{E}\Big(\frac{|X_n - c|}{1 + |X_n - c|}\Big) = 0.$$

ce qui entraîne $X_n \xrightarrow[n\to\infty]{\mathbb{P}} c = X$ par le théorème 4.2.1. □

Enfin, le théorème de Slutsky suivant s'avèrera très utile en statistique.

Théorème 4.2.6. Théorème de Slutsky. *Soit (U_n) une suite de variables aléatoires convergeant en loi et (V_n) une suite de variables aléatoires convergeant en probabilité vers une constante c. Alors pour toute fonction continue g, la suite $(g(U_n, V_n))$ a même limite en loi que la suite $(g(U_n, c))$.*

4.3 Loi forte des grands nombres

Il y a diverses versions de la loi forte des grands nombres. Nous en énonçons une.

Théorème 4.3.1 (Loi forte des grands nombres)**.** *Soit (X_i) suite de v.a.r. i.i.d. intégrables (elles sont dans L^1). On a*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\text{p.s.}} \mathbb{E}(X_1).$$

Remarque 4.3.1. Noter la différence avec la loi faible des grands nombres : le mode de convergence obtenu est plus fort, pourtant l'hypothèse est plus faible.

Démonstration. Pour la preuve du résultat dans le cas général, très longue et technique, nous renvoyons à [1] théorème V.5.2.

Ici nous allons nous faciliter la vie en faisant les hypothèses restrictives suivantes : les X_i sont de carré intégrables (dans L^2), centrées (i.e. $\mathbb{E}(X_1) = 0$) et avec $\mathbb{V}\mathrm{ar}(X_1) = 1$. En fait nous voulons surtout montrer comment on récupère la convergence presque sûre, le cas dans lequel nous nous plaçons fait déjà passer les idées générales.

Etape 1. Pour $m \in \mathbb{N}^*$ on considère $Y_m = \frac{1}{m^2}\sum_{i=1}^{m^2} X_i$. Comme dans l'exemple 4.1.2 par Tchebychev on a

$$\forall \varepsilon > 0, \quad \mathbb{P}(|Y_m| > \varepsilon) \le \frac{\mathbb{V}\mathrm{ar}(Y_m)}{\varepsilon^2} = \frac{m^2}{\varepsilon^2 m^4}\mathbb{V}\mathrm{ar}(X_1) = \frac{1}{\varepsilon^2 m^2}.$$

Comme $\sum \frac{1}{\varepsilon^2 m^2}$ est convergente la série $\sum \mathbb{P}(|Y_m| > \varepsilon)$ est aussi convergente (pour tout $\varepsilon > 0$) et par le lemme 4.1.2 on a $Y_m \xrightarrow[n\to\infty]{\text{p.s.}} 0$.

Etape 2. Pour tout n soit m tel que $m^2 \le n \le (m+1)^2 - 1$; on a

$$\begin{aligned}
\mathbb{P}\big(\big|\tfrac{1}{n}\textstyle\sum_{i=m^2+1}^{n} X_i\big| > \varepsilon\big) &= \mathbb{P}\big(\big|\textstyle\sum_{i=m^2+1}^{n} X_i\big| > n\varepsilon\big) \\
&\le \frac{n-m^2}{n^2\varepsilon^2}\mathbb{V}\mathrm{ar}(X_1) \\
&\le \frac{(m+1)^2 - 1 - m^2}{n^2\varepsilon^2} \\
&= \frac{2m}{n^2\varepsilon^2} \\
&\le \frac{2}{n^{3/2}\varepsilon^2}.
\end{aligned}$$

Par les mêmes arguments qu'à la question 1) on conclut que $\frac{1}{n}\sum_{i=m^2+1}^{n} X_i \xrightarrow[n\to\infty]{\text{p.s.}} 0$.

Etape 3. Donc finalement, en écrivant

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{m^2}{n}\frac{1}{m^2}\sum_{i=1}^{m^2} X_i + \frac{1}{n}\sum_{i=m^2+1}^{n} X_i$$

on conclut des étapes 1 et 2 que $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\text{p.s.}} 0$ car de surcroît $\frac{m^2}{n} \leq 1$. □

Nonobstant les étapes de centrage et de réduction qui peuvent être requises le résultat que nous venons de montrer permet de traiter l'exemple introductif du début de la section 4.1.

En effet les $\mathbf{1}_{F_i}$ sont i.i.d., de carré intégrable, de moyenne commune $\frac{1}{2}$. Donc

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{F_i} \xrightarrow[n\to\infty]{\text{p.s.}} \frac{1}{2}$$

ce qui nous dit bien qu'on a (4.1.1) pour presque tout ω.

4.4 Théorème central limite

Nous savons par la loi des grands nombres que la moyenne empirique (de v.a. i.i.d.) tend vers la moyenne mathématique (l'espérance commune de ces v.a.). Mais que peut-on dire de la distribution de l'écart entre moyenne empirique et espérance ?

C'est ce sur quoi nous renseigne le théorème central limite (TCL), dont l'importance est capitale en probabilités et statistiques.

Théorème 4.4.1 (TCL). *Soit (X_n) suite i.i.d. de v.a.r. dans L^2. Posons $\sigma^2 = \mathbb{V}\mathrm{ar}(X_1)$. Alors*

$$\frac{\sqrt{n}}{\sigma}\big(\overline{X}_n - \mathbb{E}(X_1)\big) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1),$$

où $\mathcal{N}(0,1)$ désigne la loi normale centrée réduite.

Démonstration. Notons $Y_n = \frac{\sqrt{n}}{\sigma}\big(\overline{X}_n - \mathbb{E}(X_1)\big) = \frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(X_i - \mathbb{E}(X_1))$ et φ la fonction caratéristique de $X_1 - \mathbb{E}(X_1)$. Par indépendance des X_i on a par le théorème 3.2.2 :

$$\begin{aligned}
\varphi_{Y_n}(t) &= \varphi_{\sum_{i=1}^{n}(X_i - \mathbb{E}(X_1))}\Big(\frac{t}{\sigma\sqrt{n}}\Big) \\
&= \prod_{i=1}^{n} \varphi\Big(\frac{t}{\sigma\sqrt{n}}\Big) \\
&= \Big(\varphi\Big(\frac{t}{\sigma\sqrt{n}}\Big)\Big)^n
\end{aligned}$$

Ensuite on utilise le théorème 3.2.4 pour voir que $\varphi'(0) = i\mathbb{E}[X_1 - \mathbb{E}(X_1)] = 0$ et

$$\varphi''(0) = -\mathbb{E}[(X_1 - \mathbb{E}(X_1))^2] = -\sigma^2.$$

Si on fait le développement de Taylor de φ en zéro à l'ordre 2 il vient :

$$\varphi(t) = 1 - \frac{\sigma^2 t^2}{2} + t^2 o(t).$$

On a donc

$$\begin{aligned}
\varphi_{Y_n}(t) &= \Big(\varphi\Big(\frac{t}{\sigma\sqrt{n}}\Big)\Big)^n \\
&= e^{n \ln \varphi(\frac{t}{\sigma\sqrt{n}})} \\
&= e^{n \ln(1 - \frac{t^2}{2n} + \frac{t^2}{n\sigma^2} o(\frac{t}{\sigma\sqrt{n}}))}
\end{aligned}$$

(où $\ln$ désigne la valeur principale du logarithme complexe ; nous ne rentrons pas dans les détails d'analyse complexe ; il faut comprendre que cette fonction vaut zéro en un et qu'on peut en faire un développement de Taylor autour de 1). En considérant que $\ln(1 - \frac{t^2}{2n} + \frac{t^2}{n\sigma^2} o(\frac{t}{\sigma\sqrt{n}})) = -\frac{t^2}{2n} + \ldots$ et en faisant $n \to \infty$ on obtient

$$\lim_{n\to\infty} \varphi_{Y_n}(t) = e^{-t^2/2}.$$

On reconnait la fonction caractéristique d'une $\mathcal{N}(0,1)$ et on conclut par le théorème 4.2.3. $\square$

Chapitre 5

Concepts de l'inférence statistique

Ce chapitre introduit la partie du cours consacrée à la statistique. Après avoir décrit la démarche statistique et fait le lien avec les probabilités, on introduit la notion de modèle statistique. Puis on rappelle les définitions d'estimateur et estimation, ainsi que les deux principales méthodes d'estimation, la méthode des moments et la méthode du maximum de vraisemblance. On s'intéresse ensuite aux critères de qualité d'un estimateur.

Pour l'ensemble de la partie statistique du cours, des livres de référence en français sont [3, 6, 7].

5.1 La démarche statistique

La **statistique** est la science dont l'objet est de recueillir, de traiter et d'analyser des **données** issues de l'observation de phénomènes **aléatoires**, c'est-à-dire dans lesquels le hasard intervient. Les méthodes statistiques se répartissent en deux classes :

- La **statistique descriptive** a pour but de **décrire** les données et de **résumer l'information** qu'elles contiennent de façon synthétique et efficace. Elle utilise pour cela des **représentations de données** sous forme de graphiques (par exemple des histogrammes ou des box-plots), de tableaux et d'indicateurs numériques (par exemple des moyennes ou des médianes empiriques). Elle permet de dégager les caractéristiques essentielles du phénomène étudié et de suggérer des hypothèses pour une étude ultérieure plus sophistiquée. Les probabilités n'ont ici qu'un rôle mineur.

- La **statistique inférentielle** va au delà de la simple description des données. Elle a pour but de **faire des prévisions** et d'aider à **prendre des décisions** au vu des observations, dans un contexte d'incertitude. En général, il faut pour cela proposer des modèles probabilistes du phénomène aléatoire étudié et savoir gérer les risques d'erreurs. Les probabilités jouent ici un rôle fondamental.

La statistique et les probabilités sont les deux aspects complémentaires de l'étude des phénomènes aléatoires. Ils sont cependant de natures différentes. Construites sur les bases théoriques du **calcul des probabilités** présentées dans les chapitres précédents, les **probabilités appliquées** proposent des **modèles probabilistes** du déroulement de phénomènes aléatoires concrets. On peut alors, **préalablement à toute expérience**, faire des prévisions sur ce qui va se produire.

Par exemple, il est usuel de modéliser la durée de bon fonctionnement ou durée de vie d'un système, mettons une ampoule électrique, par une variable aléatoire X de loi exponentielle de paramètre λ. Ayant adopté ce modèle probabiliste, on peut effectuer tous les calculs que l'on veut. Par exemple :

- La probabilité que l'ampoule ne soit pas encore tombée en panne à la date t est $\mathbb{P}(X > t) = 1 - F_X(t) = e^{-\lambda t}$, où F_X est la fonction de répartition de X.
- La durée de vie moyenne d'une ampoule est l'espérance de X : $\mathbb{E}[X] = 1/\lambda$.
- Si n ampoules identiques sont mises en fonctionnement en même temps, et qu'elles fonctionnent indépendamment les unes des autres, le nombre N_t d'ampoules qui tomberont en panne avant un instant t est une variable aléatoire de loi binomiale $\mathcal{B}\left(n, \mathbb{P}(X \leq t)\right) = \mathcal{B}\left(n, 1 - e^{-\lambda t}\right)$. Donc on s'attend à ce que, en moyenne, $\mathbb{E}[N_t] = n\left(1 - e^{-\lambda t}\right)$ ampoules tombent en panne entre 0 et t.

Dans la pratique, il est intéressant pour un utilisateur de ces ampoules d'avoir une évaluation de leur durée de vie moyenne, de la probabilité qu'elles fonctionnent correctement pendant plus d'un mois ou plus d'un an, du nombre d'ampoules qui tomberont en panne au cours d'une année, etc... Mais si l'on veut utiliser les résultats théoriques énoncés plus haut, il faut d'une part pouvoir s'assurer qu'on a choisi un bon modèle, c'est-à-dire que la durée de vie de ces ampoules est bien une variable aléatoire de loi exponentielle, et, d'autre part, pouvoir calculer d'une manière ou d'une autre la valeur du paramètre λ. C'est la statistique qui va permettre de résoudre ces problèmes. Pour cela, il faut faire une expérimentation, recueillir des données et les analyser.

On met donc en place ce qu'on appelle une **expérience** ou un essai. On fait fonctionner en parallèle et indépendamment les unes des autres $n = 10$ ampoules identiques, dans les mêmes conditions expérimentales, et on relève leurs durées de vie. Admettons que l'on obtienne les durées de vie suivantes, exprimées en heures :

91.6 35.7 251.3 24.3 5.4 67.3 170.9 9.5 118.4 57.1

Notons $x_1, \ldots, x_n$ ces observations. Il est bien évident que la durée de vie des ampoules n'est pas prévisible avec certitude à l'avance. On va donc considérer que $x_1, \ldots, x_n$ sont les **réalisations** de variables aléatoires $X_1, \ldots, X_n$. Cela signifie qu'avant l'expérience, la durée de vie de la $i^{\text{ème}}$ ampoule est inconnue et que l'on traduit cette incertitude en modélisant cette durée par une variable aléatoire X_i. Mais après l'expérience, la durée de vie a été observée. Il n'y a donc plus d'incertitude, cette durée est égale au réel x_i. On dit que x_i est la **réalisation** de X_i sur l'essai effectué.

Puisque les ampoules sont identiques, il est naturel de supposer que les X_i sont de même loi. Cela signifie qu'on observe plusieurs fois le même phénomène aléatoire. Mais le hasard fait que les réalisations de ces variables aléatoires de même loi sont différentes, d'où la variabilité dans les données. Puisque les ampoules ont fonctionné indépendamment les unes des autres, on pourra également supposer que les X_i sont des variables aléatoires indépendantes. On peut alors se poser les questions suivantes :

1. Au vu de ces observations, est-il raisonnable de supposer que la durée de vie d'une ampoule est une variable aléatoire de loi exponentielle ? Si non, quelle autre loi serait plus appropriée ? C'est un problème de **choix de modèle** ou de **test d'adéquation**.

2. Si le modèle de loi exponentielle a été retenu, comment proposer une valeur (ou un ensemble de valeurs) vraisemblable pour le paramètre λ ? C'est un problème d'**estimation paramétrique**.

3. Dans ce cas, peut-on garantir que λ est inférieur à une valeur fixée λ_0 ? Cela garantira alors que $\mathbb{E}[X] = 1/\lambda \geq 1/\lambda_0$, autrement dit que les ampoules seront suffisamment fiables. C'est un problème de **test d'hypothèses paramétriques**.

4. Sur un parc de 100 ampoules, à combien de pannes peut-on s'attendre en moins de 50 h ? C'est un problème de **prévision**.

Le premier problème central est celui de l'**estimation** : comment proposer, au vu des observations, une approximation des grandeurs inconnues du problème qui soit la plus proche possible de la réalité ? La première question peut se traiter en estimant la fonction de répartition ou la densité de la loi de probabilité sous-jacente, la seconde revient à estimer un paramètre de cette loi, la quatrième à estimer un nombre moyen de pannes sur une période donnée.

Le second problème central est celui des **tests d'hypothèses** : il s'agit de se prononcer sur la validité d'une hypothèse liée au problème : la loi est-elle exponentielle ? λ est-il inférieur à λ_0 ? un objectif de fiabilité est-il atteint ? En répondant oui ou non à ces questions, il est possible que l'on se trompe. Donc, à toute réponse statistique, il faudra associer le **degré de confiance** que l'on peut accorder à cette réponse.

En PS1, nous avons vu des méthodes permettant d'apporter des éléments de réponse aux questions ci-dessus :

1. Un **histogramme** est une estimation de densité. La densité de la loi exponentielle $\exp(\lambda)$ est $f_X(t) = \lambda e^{-\lambda t}$. Donc si un histogramme des durées de vie des 10 ampoules a une forme proche d'une fonction exponentielle décroissante, le modèle de loi exponentielle pour ces données est plausible. Inversement, si la forme d'un histogramme ne ressemble pas du tout à une exponentielle décroissante, on conclura que la loi exponentielle n'est pas un bon modèle pour la durée de vie des ampoules.

Un **graphe de probabilités** est une procédure graphique, basée sur une transformation d'une fonction de répartition cible, qui permet de juger si des données sont compatibles avec un modèle présumé. Chaque graphe est spécifique au modèle testé, donc il faut ici déterminer comment construire un graphe de probabilités pour un échantillon de loi exponentielle.

2. Si les méthodes ci-dessus amènent à admettre que la durée de vie de ces ampoules est de loi exponentielle, les procédures d'estimation par **la méthode du maximum de vraisemblance** et **la méthode des moments** permettent d'estimer le paramètre λ de cette loi. En l'occurrence, les deux méthodes donnent ici le même résultat, qui consiste à estimer λ par l'inverse de la moyenne empirique $\bar{x}_n$ des observations.

   Par ailleurs, un **intervalle de confiance** permet de donner un ensemble de valeurs vraisemblables pour λ. Si le seuil de l'intervalle de confiance est $\alpha \in [0,1]$, cela signifie qu'on a une confiance de $1-\alpha$ dans le fait que λ est dans cet intervalle.

3. On a vu en PS1 le principe des **tests d'hypothèses**, dans le cadre de tests dits bilatéraux du type test de H_0 : "$\lambda = \lambda_0$" contre H_1 : "$\lambda \neq \lambda_0$". Il s'agit ici de tester H_0 : "$\lambda \geq \lambda_0$" contre H_1 : "$\lambda < \lambda_0$".

Pour résumer, la démarche probabiliste suppose que la nature du hasard est connue. Cela signifie que l'on adopte un modèle probabiliste particulier (ici la loi exponentielle), qui permettra d'effectuer des prévisions sur les observations futures. Dans la pratique, la nature du hasard est inconnue. La statistique va, au vu des observations, formuler des hypothèses sur la nature du phénomène aléatoire étudié. Maîtriser au mieux cette incertitude permettra de traiter les données disponibles. Probabilités et statistiques agissent donc en aller-retour dans le traitement mathématique des phénomènes aléatoires.

L'exemple des ampoules est une illustration du cas le plus fréquent où les données se présentent sous la forme d'une suite de nombres. C'est ce cas que nous traiterons essentiellement dans ce cours, mais les données peuvent être beaucoup plus complexes : des fonctions, des images, etc... Les principes et méthodes généraux que nous traiterons dans ce cours sont adaptables à tous les types de données et seront mis en oeuvre dans certains cours de deuxième et troisième année.

Remarque 5.1.1. Il est très important de ne pas confondre les variables alétoires sous-jacentes X_i et leurs réalisations x_i. C'est pourquoi on adopte la convention consistant à représenter les variables aléatoires par des lettres majuscules et leurs réalisations par des lettres minuscules.

5.2 Le modèle statistique

Un modèle statistique est un objet mathématique associé à l'observation de données issues d'un phénomène aléatoire. Une **expérience statistique** consiste à recueillir une observation x d'un élément aléatoire X, à valeurs dans un espace $\mathcal{X}$ et dont on ne connait pas exactement la loi de probabilité $\mathbb{P}^X$. Des considérations de modélisation du phénomène observé amènent à admettre que $\mathbb{P}^X$ appartient à une famille $\mathcal{P}$ de lois de probabilité possibles.

Par exemple, si l'expérience est un tirage au sort selon une loi normale, l'observation peut être $x = 2.47$, réalisation d'une variable aléatoire X à valeurs dans $\mathcal{X} = \mathbb{R}$, où X est de loi normale $\mathbb{P}^X = \mathcal{N}(m_0, \sigma_0^2)$ appartenant à la famille des lois normales $\mathcal{P} = \{\mathcal{N}(m, \sigma^2); m \in \mathbb{R}, \sigma^2 \in \mathbb{R}^{+*}\}$.

L'élément aléatoire X est défini sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ et à valeurs dans $\mathcal{X}$ muni d'une tribu $\mathcal{A}$. Autrement dit, c'est une application mesurable de $(\Omega, \mathcal{F})$ dans $(\mathcal{X}, \mathcal{A})$. La loi de probabilité $\mathbb{P}^X$ de X est déterminée par $\forall A \in \mathcal{A}, \mathbb{P}^X(A) = \mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A))$.

Définition 5.2.1. : Le **modèle statistique** (ou la structure statistique) associé à une expérience statistique est le triplet $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, où :

- $\mathcal{X}$ est l'espace des observations, ensemble de toutes les valeurs possibles pour X.
- $\mathcal{A}$ est la tribu des évènements observables associée.
- $\mathcal{P}$ est la famille des lois de probabilités possibles pour X, définie sur $\mathcal{A}$.

L'intérêt de cette notion de modèle statistique est qu'elle permet de traiter avec le même formalisme tous les types d'observations possibles.

- On dit que le modèle est **discret** quand $\mathcal{X}$ est fini ou dénombrable. Dans ce cas, la tribu $\mathcal{A}$ est l'ensemble des parties de $\mathcal{X}$: $\mathcal{A} = \mathcal{P}(\mathcal{X})$. C'est le cas quand l'élément aléatoire observé X a une loi de probabilité discrète.
- On dit que le modèle est **continu** quand $\mathcal{X} \subset \mathbb{R}^p$ et $\forall \, \mathbb{P}^X \in \mathcal{P}$, $\mathbb{P}^X$ admet une densité (par rapport à la mesure de Lebesgue) dans $\mathbb{R}^p$. Dans ce cas, $\mathcal{A}$ est la tribu des boréliens de $\mathcal{X}$: $\mathcal{A} = \mathcal{B}(\mathcal{X})$.
- On peut aussi envisager des modèles ni continus ni discrets, par exemple si l'observation a certains éléments continus et d'autres discrets. $\mathcal{X}$ et $\mathcal{A}$ sont alors plus complexes.

Le cas le plus fréquent est celui où l'élément aléatoire observé est un vecteur de variables aléatoires indépendantes et de même loi (i.i.d.) : $X = (X_1, \ldots, X_n)$, où les X_i sont i.i.d. On dit que l'on a alors un **modèle d'échantillon**. Dans ce cas, par convention, si on note $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ le modèle correspondant à un échantillon de taille 1, on notera $(\mathcal{X}, \mathcal{A}, \mathcal{P})^n$ le modèle correspondant à un échantillon de taille n.

Exemple 5.2.1. *Exemple des ampoules.* Reprenons l'expérience vue en Section 5.1 consistant à recueillir les durées de vie, supposées indépendantes et de même loi exponentielle, de n ampoules électriques. L'observation est de la forme $x = (x_1, \ldots, x_n)$, où les x_i sont des réalisations de variables aléatoires X_i indépendantes et de même loi exponentielle de paramètre λ inconnu.

Pour tout i, $x_i \in \mathbb{R}^+$, donc l'espace des observations est $\mathcal{X} = \mathbb{R}^{+^n}$. Alors la tribu associée est $\mathcal{A} = \mathcal{B}(\mathbb{R}^{+^n})$. Le modèle est continu. Comme on admet que la loi est exponentielle mais que son paramètre est inconnu, l'ensemble des lois de probabilités possibles pour chaque X_i est $\{\exp(\lambda); \lambda > 0\}$. Comme les X_i sont indépendantes, si elles sont de loi $\exp(\lambda)$, la loi de probabilité du vecteur $X = (X_1, \ldots, X_n)$ est la loi produit $\mathbb{P}^X = \exp(\lambda)^{\otimes n}$, loi d'un vecteur aléatoire de taille n dont les composantes sont indépendantes et de même loi $\exp(\lambda)$.

Finalement, le modèle statistique associé est :

$$\left(\mathbb{R}^{+^n}, \mathcal{B}(\mathbb{R}^{+^n}), \left\{\exp(\lambda)^{\otimes n}; \lambda > 0\right\}\right)$$

qu'on peut aussi écrire, d'après la convention énoncée :

$$\left(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+), \{\exp(\lambda); \lambda > 0\}\right)^n.$$

Exemple 5.2.2. *Exemple du contrôle de qualité.* Une chaîne de production produit un très grand nombre de pièces et on s'intéresse à la proportion inconnue de pièces défectueuses. Pour l'estimer, on prélève indépendamment n pièces dans la production et on les contrôle. L'observation est $x = (x_1, \ldots, x_n)$, où :

$$x_i = \begin{cases} 1 \text{ si la } i^{\text{ème}} \text{ pièce est défectueuse} \\ 0 \text{ sinon} \end{cases}$$

Par conséquent, l'espace des observations est $\mathcal{X} = \{0, 1\}^n$. Il est fini, donc le modèle est discret et $\mathcal{A} = \mathcal{P}(\{0, 1\}^n)$. Les X_i sont indépendantes et de même loi de Bernoulli $\mathcal{B}(p)$, où $p = \mathbb{P}(X_i = 1)$ est la probabilité qu'une pièce soit défectueuse.

Alors le modèle statistique peut s'écrire :

$$\left(\{0, 1\}^n, \mathcal{P}(\{0, 1\}^n), \left\{\mathcal{B}(p)^{\otimes n}; p \in [0, 1]\right\}\right)$$

ou

$$\left(\{0, 1\}, \mathcal{P}(\{0, 1\}), \{\mathcal{B}(p); p \in [0, 1]\}\right)^n.$$

Remarque 5.2.1. Quand l'élément aléatoire X est numérique, il admet une fonction de répartition F. D'après le théorème 2.3.1, la fonction de répartition caractérise une loi de probabilité, donc l'ensemble $\mathcal{P}$ des lois de probabilité possibles pour X est en bijection avec l'ensemble $\mathcal{F}$ des fonctions de répartition possibles. Aussi le modèle statistique peut dans ce cas être noté $(\mathcal{X}, \mathcal{A}, \mathcal{F})$ au lieu de $(\mathcal{X}, \mathcal{A}, \mathcal{P})$.

Définition 5.2.2. Dans un modèle statistique $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, une **statistique** est une application mesurable t de $(\mathcal{X}, \mathcal{A})$ dans un espace $\mathcal{Y}$ muni d'une tribu $\mathcal{B}$.

Concrètement, cela signifie qu'une statistique est une fonction t des observations x telle que l'on peut calculer la probabilité de tout évènement de la forme $[t(X) \in B]$, $\forall\, B \in \mathcal{B}$. En particulier, t ne doit pas dépendre de paramètres inconnus.

Puisque x est une réalisation de l'élément aléatoire X, $t(x)$ est une réalisation de l'élément aléatoire $T = t(X)$.

5.3 Modèle paramétrique ou non paramétrique

Un **modèle paramétrique** est un modèle où l'on suppose que le type de loi de X est connu, mais qu'il dépend d'un paramètre θ inconnu, de dimension d. Alors, la famille de lois de probabilité possibles pour X peut s'écrire $\mathcal{P} = \left\{\mathbb{P}_\theta^X \,;\, \theta \in \Theta \subset \mathbb{R}^d\right\}$, où Θ est l'ensemble des valeurs possibles pour θ.

C'est évidemment le cas des deux exemples. Le problème principal est alors de faire de l'inférence statistique sur θ : l'estimer, ponctuellement ou par régions de confiance (intervalles de confiance si $d = 1$), et effectuer des tests d'hypothèses portant sur θ. On fait alors de la **statistique paramétrique**.

Dans un modèle paramétrique, pour bien mettre en évidence le paramètre θ, on le fera figurer explicitement dans les fonctions de base comme les probabilités élémentaires $\mathbb{P}(X_i = x_i; \theta)$, les fonctions de répartition $F_{X_i}(x_i; \theta)$ et les densités $f_{X_i}(x_i; \theta)$.

Un **modèle non paramétrique** est un modèle où $\mathcal{P}$ ne peut pas se mettre sous la forme ci-dessus. Par exemple, $\mathcal{P}$ peut être :

- l'ensemble des lois de probabilité continues sur $\mathbb{R}$,
- l'ensemble des lois de probabilité dont le support est $[0, 1]$,
- l'ensemble des lois de probabilité sur $\mathbb{R}$ symétriques par rapport à l'origine,
- etc...

Dans ce cadre, il est possible de calculer des estimations, des intervalles de confiance, d'effectuer des tests d'hypothèses. Mais les objets sur lesquels portent ces procédures statistiques ne sont plus des paramètres de lois de probabilité. On peut vouloir estimer des quantités réelles comme l'espérance et la variance des observations. Par exemple, on a vu en PS1 qu'on pouvait utiliser la moyenne empirique des données pour estimer l'espérance de la loi sous-jacente, sans faire d'hypothèses sur cette loi. On peut aussi vouloir estimer des fonctions, comme la fonction de répartition et la densité des observations. On a vu en PS1 qu'un histogramme est une estimation de densité.

En termes de tests d'hypothèses, on peut effectuer des tests sur la valeur d'une espérance, tester si les observations sont indépendantes, si elles présentent une croissance, si elles proviennent d'une loi normale, tester si plusieurs échantillons proviennent de la même loi, etc... On fait alors de la **statistique non paramétrique**.

De manière générale, la statistique non paramétrique regroupe l'ensemble des méthodes statistiques qui permettent de tirer de l'information pertinente de données sans faire l'hypothèse que la loi de probabilité de ces observations appartient à une famille paramétrée connue.

Un des problèmes de la statistique paramétrique est le risque d'erreur du à un mauvais choix de modèle. Par exemple, si on effectue des calculs en supposant que des observations sont de loi exponentielle alors qu'en fait elles sont de loi normale, on obtiendra des résultats aberrants. L'avantage de la statistique non paramétrique est de ne pas être soumise à cet aléa (on dit qu'elle est *robuste*). En revanche, si les observations sont bien issues d'un modèle précis, les méthodes statistiques paramétriques qui utilisent ce modèle seront plus performantes que celles qui ne l'utilisent pas. Il est donc également important d'établir des méthodes permettant de déterminer si des observations sont issues ou non de tel ou tel modèle paramétrique, les *tests d'adéquation*.

5.4 Estimateur et méthodes d'estimation

Dans le modèle paramétrique $(\mathcal{X}, \mathcal{A}, \left\{\mathbb{P}_\theta^X \,;\, \theta \in \Theta \subset \mathbb{R}^d\right\})$, on observe la réalisation x d'un élément aléatoire X de loi de probabilité $\mathbb{P}_\theta^X$ dépendant d'un paramètre θ inconnu de dimension d. Estimer θ consiste à donner, au vu de l'observation x, une approximation ou une évaluation de θ que l'on espère la plus proche possible de la vraie valeur inconnue.

La plupart du temps, on se placera dans un modèle d'échantillon. Mais la majeure partie des résultats qui suivent peuvent s'exprimer pour un modèle paramétrique quelconque.

5.4.1 Définition d'un estimateur

Pour estimer θ on ne dispose que de l'observation x, donc une estimation de θ sera une fonction de ces observations.

Définition 5.4.1. Un **estimateur** de θ est une statistique t à valeurs dans l'ensemble Θ des valeurs possibles de θ. Autrement dit, c'est une application mesurable de $(\mathcal{X}, \mathcal{A})$ dans Θ muni d'une certaine tribu $\mathcal{B}$. Très souvent, $(\Theta, \mathcal{B}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.

$$t : \begin{array}{l} (\mathcal{X}, \mathcal{A}) \to (\Theta, \mathcal{B}) \\ x \mapsto t(x) \end{array}$$

Une convention commode que nous adopterons est d'appeler également **estimateur** de θ l'élément aléatoire $T = t(X)$. Sa réalisation $t(x)$ est appelée **estimation** de θ.

Remarque 5.4.1. Un estimateur est donc une quantité aléatoire, alors qu'une estimation est une quantité déterministe, calculée à partir des observations. **Il ne faut pas confondre un paramètre et son estimation, ni une estimation et un estimateur**.

Dans l'exemple des ampoules, on a un échantillon de loi exponentielle de paramètre λ. λ une constante réelle positive inconnue. Comme vu en PS1, la méthode du maximum de vraisemblance nous amène à estimer λ par l'inverse de la moyenne empirique des observations. L'estimation de λ, notée usuellement $\hat{\lambda}_n$, est donc $\hat{\lambda}_n = 1/\bar{x}_n = n/\sum_{i=1}^n x_i$. Dans l'exemple, $\bar{x}_n$=83.15, donc $\hat{\lambda}_n = 1/83.15 = 0.012$. C'est une valeur numérique calculée à partir des données, qui est une valeur vraisemblable pour λ.

L'estimateur de λ est la variable aléatoire $1/\bar{X}_n = n/\sum_{i=1}^n X_i$ dont $\hat{\lambda}_n$ est la réalisation pour ces données. On a dit dans la section 5.1 que l'on notait les variables aléatoires à l'aide de majuscules (ex : X_i) et leurs réalisations à l'aide de minuscules (ex : x_i). Cette règle voudrait donc que l'on note $\hat{\Lambda}_n = 1/\bar{X}_n$ l'estimateur de λ. Malheureusement, l'utilisation des lettres grecques majuscules est peu répandue et l'usage veut qu'on utilise la même notation $\hat{\lambda}_n$ pour l'estimation $1/\bar{x}_n$ et l'estimateur $1/\bar{X}_n$. Nous adopterons donc cette convention, le contexte permettant de déterminer si, quand on utilise $\hat{\lambda}_n$, on parle de l'estimation ou de l'estimateur. Néanmoins, il est important de ne pas confondre ces deux notions.

De manière générale, un estimateur peut être compris comme une méthode d'estimation (ici prendre l'inverse de la moyenne empirique) et l'estimation est la valeur numérique obtenue par cette méthode sur les données étudiées.

Dans un modèle paramétrique, la quantité à estimer n'est pas forcément le paramètre θ lui même, mais est forcément une fonction de θ, $\varphi(\theta)$. Alors, si T est un estimateur de θ, on estimera $\varphi(\theta)$ par $\varphi(T)$.

Pour un modèle d'échantillon de taille n, l'estimateur est traditionnellement noté T_n et l'estimation t_n.

Dans les sections qui suivent, on présente les plus usuelles des méthodes d'estimation, la méthode des moments et la méthode du maximum de vraisemblance.

5.4.2 La méthode des moments

Intuitivement, il parait logique d'estimer une espérance par une moyenne empirique, une variance par une variance empirique, etc... Plus généralement, le principe de la méthode des moments est d'estimer un moment théorique par le moment empirique de même ordre.

On a vu dans le chapitre 2 que le **moment d'ordre** k de la loi des X_i est $m_k = \mathbb{E}[X_1^k]$. De même, on définit le **moment centré d'ordre** k par $\mu_k = \mathbb{E}\left[(X_1 - \mathbb{E}[X_1])^k\right]$. L'espérance est donc le moment d'ordre 1 et la variance est le moment centré d'ordre 2. Rappelons que, d'après le théorème 3.2.4, un moyen commode de calculer tous les moments est d'utiliser la fonction caractéristique de la loi des X_i.

On considère un modèle d'échantillon $X = (X_1, \ldots, X_n)$, où les X_i sont des variables aléatoires réelles indépendantes et de même loi. Le **moment empirique d'ordre** k de l'échantillon est $m_k^e = \dfrac{1}{n}\displaystyle\sum_{i=1}^n X_i^k$ et le **moment empirique centré d'ordre** k de l'échantillon est $\mu_k^e = \dfrac{1}{n}\displaystyle\sum_{i=1}^n (X_i - \bar{X}_n)^k$.

Définition 5.4.2. L'**estimateur** de m_k **par la méthode des moments (EMM)** est m_k^e. L'EMM de μ_k est μ_k^e. Si le paramètre à estimer est θ, l'EMM de θ est usuellement noté $\tilde{\theta}_n$.

Si le paramètre à estimer n'est pas un moment de la loi des X_i, il suffit d'adapter le principe précédent pour obtenir l'EMM. Par exemple, l'EMM de l'espérance $\mathbb{E}[X_1]$ est la moyenne empirique $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Si le paramètre à estimer est $\theta \in \mathbb{R}$ et que $\mathbb{E}[X_1] = \varphi(\theta)$, où φ est une fonction inversible, alors l'EMM de θ est $\tilde{\theta}_n = \varphi^{-1}\left(\bar{X}_n\right)$.

L'EMM de $\mathbb{V}\mathrm{ar}[X_1]$ est la variance empirique de l'échantillon $S_n^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2$.

Plus généralement, si la loi des X_i a deux paramètres réels θ_1 et θ_2 tels que $(\mathbb{E}[X_1], \mathbb{V}\mathrm{ar}[X_1]) = \varphi\left(\theta_1, \theta_2\right)$, où φ est une fonction inversible, alors les estimateurs de θ_1 et θ_2 par la méthode des moments sont $\left(\tilde{\theta}_{1n}, \tilde{\theta}_{2n}\right) = \varphi^{-1}\left(\bar{X}_n, S_n^2\right)$.

Exemple 5.4.1. *loi de Bernoulli* (exemple du contrôle de qualité). Si $X_1, \ldots, X_n$ sont indépendantes et de même loi de Bernoulli $\mathcal{B}(p)$, $\mathbb{E}[X_1] = p$. Donc l'EMM de p est $\tilde{p}_n = \bar{X}_n$. Cet estimateur n'est autre que la proportion de 1 dans l'échantillon. Ce résultat signifie que l'on peut estimer la probabilité qu'un évènement survienne par le pourcentage de fois où cet évènement est survenu dans une suite d'expériences identiques et indépendantes, ce qui conforte l'intuition.

Exemple 5.4.2. *loi exponentielle* (exemple des ampoules). Si $X_1, \ldots, X_n$ sont indépendantes et de même loi exponentielle $\exp(\lambda)$, $\mathbb{E}[X_1] = 1/\lambda$. Donc l'EMM de λ est $\tilde{\lambda}_n = 1/\bar{X}_n$.

Exemple 5.4.3. *loi normale*. Si $X_1, \ldots, X_n$ sont indépendantes et de même loi normale $\mathcal{N}\left(m, \sigma^2\right)$, $\mathbb{E}[X_1] = m$ et $\mathbb{V}\mathrm{ar}[X_1] = \sigma^2$, donc les EMM de m et σ^2 sont $\tilde{m}_n = \bar{X}_n$ et $\tilde{\sigma}_n^2 = S_n^2$.

Exemple 5.4.4. *loi gamma*. Si $X_1, \ldots, X_n$ sont indépendantes et de même loi gamma $G(a, \lambda)$, $\mathbb{E}[X_1] = a/\lambda$ et $\mathbb{V}\mathrm{ar}[X_1] = a/\lambda^2$. On en déduit facilement que :

$$\lambda = \frac{\mathbb{E}[X_1]}{\mathbb{V}\mathrm{ar}[X_1]} \quad \text{et} \quad a = \frac{\mathbb{E}[X_1]^2}{\mathbb{V}\mathrm{ar}[X_1]}$$

Donc les EMM de a et λ sont :

$$\tilde{\lambda}_n = \frac{\bar{X}_n}{S_n^2} \quad \text{et} \quad \tilde{a}_n = \frac{\bar{X}_n^2}{S_n^2}$$

5.4.3 La méthode du maximum de vraisemblance

La fonction de vraisemblance a été vue en PS1 pour des modèles d'échantillon discrets ou continus. On peut en fait en donner une définition plus générale, valable pour n'importe quel modèle paramétrique.

Définition 5.4.3. Une mesure μ sur $(\mathcal{X}, \mathcal{A})$ est dite **dominante** pour le modèle paramétrique $(\mathcal{X}, \mathcal{A}, \{\mathbb{P}_\theta^X ; \theta \in \Theta\})$ si et seulement si pour tout θ, $\mathbb{P}_\theta^X$ est absolument continue par rapport à μ ($\mathbb{P}_\theta^X << \mu$), c'est-à-dire (voir théorème **??**) :

$$\forall A \in \mathcal{A}, \mu(A) = 0 \Rightarrow \forall \theta \in \Theta, \mathbb{P}_\theta^X(A) = 0.$$

Alors le théorème **??** de Radon-Nikodym assure que $\mathbb{P}_\theta^X$ admet une densité par rapport à μ. Cette densité est appelée **fonction de vraisemblance** du modèle et est notée $\mathcal{L}(\theta; x) = \frac{d\mathbb{P}^X}{d\mu}(x)$.

La fonction de vraisemblance vérifie :

$$\forall A \in \mathcal{A},\ \mathbb{P}_\theta^X(A) = \mathbb{P}(X \in A; \theta) = \int_A \mathcal{L}(\theta; x)\, d\mu(x) \tag{5.4.1}$$

De manière équivalente, si $\mathcal{L}(\theta; x)$ vérifie (5.4.1), alors $\mathcal{L}(\theta; x)$ est la fonction de vraisemblance du modèle.

L'avantage de cette définition générale est qu'elle permet de définir la fonction de vraisemblance pour des modèles atypiques, pas forcément d'échantillon et pas forcément discrets ou continus. Mais bien entendu, pour les modèles discrets ou continus, on retrouve bien les résultats attendus.

Modèles discrets. La mesure dominante est la mesure de dénombrement sur $\mathcal{X}$ μ_d, qui vérifie $\mu_d(A) = card(A)$ et $\int_A f(x)\, d\mu_d(x) = \sum_{x\in A} f(x)$. Si X est un vecteur aléatoire de loi discrète, définie par les probabilités élémentaires $\mathbb{P}(X = x; \theta)$, alors :

$$\mathbb{P}(X \in A; \theta) = \sum_{x\in A} \mathbb{P}(X = x; \theta) = \int_A \mathbb{P}(X = x; \theta)\, d\mu_d(x).$$

Donc la fonction de vraisemblance est $\mathcal{L}(\theta; x) = \mathbb{P}(X = x; \theta)$.

Modèles continus. La mesure dominante est la mesure de Lebesgue λ_L, qui, dans $\mathbb{R}$, vérifie $\lambda_L(]a, b]) = b - a$ et $\int_{]a,b]} f(x)\, d\lambda_L(x) = \int_a^b f(x)\, dx$. Si X est un vecteur aléatoire admettant une densité $f_X(x; \theta)$ (par rapport à la mesure de Lebesgue), on a :

$$\mathbb{P}(X \in A; \theta) = \int_A f_X(x; \theta)\, dx = \int_A f_X(x; \theta)\, d\lambda_L(x).$$

Donc la fonction de vraisemblance est $\mathcal{L}(\theta; x) = f_X(x; \theta)$.

On retrouve ainsi les résultats vus en PS1 pour un modèle d'échantillon discret :

$$\mathcal{L}(\theta; x) = \mathbb{P}(X = x; \theta) = \mathcal{L}(\theta; x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n; \theta) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i; \theta)$$

et pour un modèle d'échantillon continu où la densité des observations est f :

$$\mathcal{L}(\theta; x) = f_X(x; \theta) = \mathcal{L}(\theta; x_1, \ldots, x_n) = f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f_{X_i}(x_i; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

Remarque 5.4.2. La probabilité et la densité utilisées dans ces définitions sont des fonctions des observations $x_1, \ldots, x_n$, dépendant du paramètre θ. A l'inverse, la fonction de vraisemblance est considérée comme une fonction de θ dépendant des observations $x_1, \ldots, x_n$, ce qui permet, par exemple, de dériver cette fonction par rapport à θ.

Le principe de la méthode du maximum de vraisemblance est d'estimer le paramètre θ du modèle par la valeur qui maximise la fonction de vraisemblance.

Définition 5.4.4. Soit $(\mathcal{X}, \mathcal{A}, \{\mathbb{P}_\theta^X\, ;\, \theta \in \Theta\})$ un modèle statistique paramétrique. Si la fonction de vraisemblance admet un maximum unique au point $\hat{\theta}(x)$, alors l'application $x \mapsto \hat{\theta}(x)$ est appelée **statistique de maximum de vraisemblance**.

$$\hat{\theta}(x) = \arg\max_{\theta} \mathcal{L}(\theta; x)$$

$\hat{\theta}(X)$ est l'**estimateur de maximum de vraisemblance (EMV)** de θ au vu de X.

Pour simplifier, l'estimateur $\hat{\theta}(X)$ et l'estimation $\hat{\theta}(x)$ sont usuellement plus simplement notés $\hat{\theta}$. Pour un modèle d'échantillon de taille n, on note $\hat{\theta}_n$.

Afin de comprendre pourquoi maximiser la fonction de vraisemblance permet d'obtenir une estimation de θ, on considère l'exemple élémentaire suivant.

Exemple 5.4.5. Dans cet exemple, $n = 1$. On considère que l'on sait que X_1 est de loi binomiale $\mathcal{B}(15, p)$, avec p inconnu. On observe $x_1 = 5$ et on cherche à estimer p. La fonction de vraisemblance est :

$$\mathcal{L}(p; 5) = \mathbb{P}(X_1 = 5; p) = \binom{15}{5}\, p^5\, (1-p)^{15-5}$$

C'est la probabilité d'avoir observé un 5 quand la valeur du paramètre est p. Calculons-là pour quelques valeurs de p.

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\mathcal{L}(p; 5)$	0.01	0.10	0.21	0.19	0.09	0.02	0.003	10^{-4}	$2\,10^{-7}$

On tire de cette table que quand $p = 0.8$, c'est-à-dire quand X_1 est de loi $\mathcal{B}(15, 0.8)$, il n'y a qu'une chance sur 10000 d'observer $x_1 = 5$. En revanche, il y a 21% de chances d'observer un 5 quand $p = 0.3$. Il est donc beaucoup plus vraisemblable que p soit égal à 0.3 plutôt qu'à 0.8. En suivant ce raisonnement, on aboutit à dire que la valeur la plus vraisemblable de p est celle pour laquelle la probabilité d'observer un 5 est maximale. C'est donc la valeur de p qui maximise la fonction de vraisemblance.

Pour la calculer, on peut annuler la dérivée de la vraisemblance. Mais on remarque que la vraisemblance est un produit. Comme il est plus commode de maximiser (ou de dériver) une somme qu'un produit, on utilise le fait que la valeur qui rend maximale une fonction rend aussi maximal son logarithme. On va donc plutôt maximiser le logarithme de la fonction de vraisemblance, qu'on appelle la **log-vraisemblance**. Pour notre exemple, la log-vraisemblance vaut :

$$\ln \mathcal{L}(p; x_1) = \ln \binom{15}{x_1} + x_1 \ln p + (15 - x_1) \ln(1-p)$$

Sa dérivée est :

$$\frac{\partial}{\partial p} \ln \mathcal{L}(p; x_1) = \frac{x_1}{p} - \frac{15 - x_1}{1-p} = \frac{x_1 - 15p}{p(1-p)}$$

qui s'annule pour $p = \frac{x_1}{15} = \frac{5}{15} = \frac{1}{3}$. Donc la valeur la plus vraisemblable de p est $\frac{1}{3}$. La vraisemblance maximale est $\mathcal{L}(\frac{1}{3}; 5) = 21.4\%$.

En suivant le raisonnement précédent, pour un modèle paramétrique quelconque où l'observation est x, il est logique de dire que la valeur la plus vraisemblable de θ est la valeur pour laquelle la probabilité d'observer x est la plus forte possible. Cela revient à faire comme si c'était l'éventualité la plus probable qui s'était produite au cours de l'expérience.

Comme dans l'exemple, dans la plupart des cas, la fonction de vraisemblance s'exprime comme un produit. Donc $\hat{\theta}$ sera en général calculé en maximisant la log-vraisemblance :

$$\hat{\theta} = \arg\max_{\theta} \ln \mathcal{L}(\theta; x)$$

Quand $\theta = (\theta_1, \ldots, \theta_d)^T \in \mathbb{R}^d$ et que toutes les dérivées partielles ci-dessous existent, $\hat{\theta}$ est solution du système d'équations appelées **équations de vraisemblance** :

$$\forall j \in \{1, \ldots, d\}, \quad \frac{\partial}{\partial \theta_j} \ln \mathcal{L}(\theta; x) = 0$$

A priori, une solution de ce système d'équations pourrait être un minimum de la vraisemblance. Mais on peut montrer que la nature d'une fonction de vraisemblance fait que c'est bien un maximum que l'on obtient. Ce système peut avoir plusieurs solutions s'il existe des maxima locaux. Il faudra alors déterminer la valeur de θ correspondant au maximum global.

Il est fréquent que le système des équations de vraisemblance n'ait pas de solution explicite. Dans ce cas, on le résoud par des méthodes numériques, comme les méthodes de Newton-Raphson et de Nelder-Mead.

En PS1, on a vu les résultats suivants pour des modèles d'échantillon.

Exemple 5.4.6. *loi de Bernoulli* (exemple du contrôle de qualité). Si les X_i sont de loi $\mathcal{B}(p)$, l'EMV de p est $\hat{p}_n = \bar{X}_n$.

Exemple 5.4.7. *loi exponentielle* (exemple des ampoules). Si les X_i sont de loi $\exp(\lambda)$, l'EMV de λ est $\hat{\lambda}_n = \dfrac{1}{\bar{X}_n}$.

Voyons maintenant d'autres exemples pour des modèles d'échantillon.

Exemple 5.4.8. *loi normale.* Si les X_i sont de loi $\mathcal{N}\left(m, \sigma^2\right)$, la fonction de vraisemblance est :

$$\mathcal{L}(m, \sigma^2; x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}\left(x_i; m, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\dfrac{(x_i - m)^2}{2\sigma^2}} = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n}\, e^{-\dfrac{1}{2\sigma^2}\sum\limits_{i=1}^{n}(x_i - m)^2}$$

D'où $\ln \mathcal{L}(m, \sigma^2; x_1, \ldots, x_n) = -\dfrac{n}{2}\ln \sigma^2 - \dfrac{n}{2}\ln 2\pi - \dfrac{1}{2\sigma^2}\sum\limits_{i=1}^{n}(x_i - m)^2$.

On doit annuler les dérivées partielles de ce logarithme par rapport à m et σ^2. On a :

- $\dfrac{\partial}{\partial m}\ln \mathcal{L}(m, \sigma^2; x_1, \ldots, x_n) = -\dfrac{1}{2\sigma^2}\sum\limits_{i=1}^{n} -2(x_i - m) = \dfrac{1}{\sigma^2}\left(\sum\limits_{i=1}^{n} x_i - nm\right)$, qui s'annule pour $m = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i = \bar{x}_n$.
- $\dfrac{\partial}{\partial \sigma^2}\ln \mathcal{L}(m, \sigma^2; x_1, \ldots, x_n) = -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\sum\limits_{i=1}^{n}(x_i - m)^2$, qui s'annule pour $\sigma^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n}(x_i - m)^2$.

$\hat{m}_n$ et $\hat{\sigma}_n^2$ sont les valeurs de m et σ^2 qui vérifient les deux conditions en même temps. Par conséquent, les EMV de m et σ^2 sont $\hat{m}_n = \bar{X}_n$ et $\hat{\sigma}_n^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 = S_n^2$.

Exemple 5.4.9. *loi gamma.* Si les X_i sont de loi gamma $G(a, \lambda)$, la fonction de vraisemblance est :

$$\mathcal{L}(a, \lambda; x_1, \ldots, x_n) \;=\; \prod_{i=1}^{n} f_{X_i}\left(x_i; a, \lambda\right) = \prod_{i=1}^{n} \frac{\lambda^a}{\Gamma(a)}\, e^{-\lambda x_i}\, {x_i}^{a-1} = \frac{\lambda^{na}}{\left[\Gamma(a)\right]^n}\, e^{-\lambda\sum\limits_{i=1}^{n} x_i} \prod_{i=1}^{n} {x_i}^{a-1}$$

D'où $\ln \mathcal{L}(a, \lambda; x_1, \ldots, x_n) = na\ln\lambda - n\ln\Gamma(a) - \lambda\sum\limits_{i=1}^{n} x_i + (a-1)\sum\limits_{i=1}^{n}\ln x_i$.

On doit annuler les dérivées partielles de ce logarithme par rapport à a et λ. On a :

- $\dfrac{\partial}{\partial \lambda}\ln \mathcal{L}(a, \lambda; x_1, \ldots, x_n) = \dfrac{na}{\lambda} - \sum\limits_{i=1}^{n} x_i$ qui s'annule pour $\lambda = \dfrac{na}{\sum\limits_{i=1}^{n} x_i} = \dfrac{a}{\bar{x}_n}$.
- $\dfrac{\partial}{\partial a}\ln \mathcal{L}(a, \lambda; x_1, \ldots, x_n) = n\ln\lambda - n\dfrac{\Gamma'(a)}{\Gamma(a)} + \sum\limits_{i=1}^{n}\ln x_i$.

En remplaçant λ par $a/\bar{x}_n$, on obtient que $\hat{a}_n$ est solution de l'équation implicite :

$$n\ln a - n\ln\bar{X}_n - n\frac{\Gamma'(a)}{\Gamma(a)} + \sum_{i=1}^{n}\ln X_i = 0$$

Il n'y a pas d'expression explicite de $\hat{a}_n$. Cette équation est à résoudre par des méthodes numériques. Une fois $\hat{a}_n$ déterminé, on en déduit $\hat{\lambda}_n = \hat{a}_n/\bar{X}_n$.

Remarque 5.4.3. Dans les trois premiers exemples, la méthode des moments et la méthode du maximum de vraisemblance donnent les mêmes résultats. Ce n'est le cas que pour quelques lois de probabilité parmi les plus élémentaires. En fait, dans la plupart des cas, les deux méthodes fournissent des estimateurs différents.

C'est le cas de la loi gamma. On a deux estimateurs différents pour chaque paramètre. On doit donc se demander quel est le meilleur d'entre eux. Cela amène à se poser la question de la qualité d'un estimateur, ce qui fait l'objet de la section suivante.

Remarque 5.4.4. On pourrait croire au vu des exemples que le calcul des estimateurs des moments est beaucoup plus simple que celui des estimateurs de maximum de vraisemblance. Mais ce n'est pas vrai en général.

Proposition 5.4.1. *Pour n'importe quelle fonction φ, si $\hat{\theta}_n$ est l'estimateur de maximum de vraisemblance de θ, $\varphi(\hat{\theta}_n)$ est l'estimateur de maximum de vraisemblance de $\varphi(\theta)$.*

5.5 Qualité d'un estimateur

En toute généralité, θ peut-être un paramètre à plusieurs dimensions, mais on supposera dans cette section que θ est un réel. Cela signifie par exemple que, quand on a un échantillon de loi normale $\mathcal{N}(m, \sigma^2)$, on s'intéressera séparément à la qualité des estimateurs de m et de σ^2. Les estimateurs T considérés ici seront donc des variables aléatoires réelles. Pour $\theta \in \mathbb{R}^d$, $d \geq 2$, toutes les notions de cette section seront généralisées dans le chapitre 6.

Un estimateur T de θ sera un bon estimateur s'il est suffisamment proche, en un certain sens, de θ. Il faut donc définir une mesure de l'écart entre θ et T. On appelle cette mesure le **risque** de l'estimateur. On a intérêt à ce que le risque d'un estimateur soit le plus petit possible. Par exemple, les risques suivants expriment bien un écart entre T et θ.

Définition 5.5.1.

- L'**erreur d'estimation** de T est $T - \theta$.
- L'**erreur absolue d'estimation** est $|T - \theta|$.
- L'**erreur quadratique d'estimation** est $(T - \theta)^2$.

Mais comme il est plus facile de manipuler des quantités déterministes que des quantités aléatoires, on s'intéresse en priorité aux espérances des quantités précédentes. En particulier :

Définition 5.5.2.

- Le **biais** de T est l'erreur moyenne d'estimation : $\mathbb{E}[T - \theta] = \mathbb{E}[T] - \theta$.
- Le **risque quadratique** ou **erreur quadratique moyenne (EQM)** (en anglais : MSE pour Mean Squared Error) est :

$$EQM[T] = \mathbb{E}\left[(T - \theta)^2\right] = ||T - \theta||_2^2$$

Dans le cas du biais, le risque peut être nul.

Définition 5.5.3. Un **estimateur** T de θ est **sans biais** si et seulement si $\mathbb{E}[T] = \theta$. Il est **biaisé** si et seulement si $\mathbb{E}[T] \neq \theta$.

Le biais mesure une erreur systématique d'estimation de θ par T. Par exemple, si $\mathbb{E}[T] - \theta < 0$, cela signifie que T aura tendance à sous-estimer θ. Si T est sans biais, cela signifie que l'estimateur ne va, en moyenne, ni sur-estimer ni sous-estimer θ, ce qui est souhaitable.

L'erreur quadratique moyenne s'écrit :

$$\begin{aligned}
EQM[T] &= \mathbb{E}\left[(T-\theta)^2\right] = \mathbb{E}\left[(T-\mathbb{E}[T]+\mathbb{E}[T]-\theta)^2\right] \\
&= \mathbb{E}\left[(T-\mathbb{E}[T])^2\right] + 2\mathbb{E}\left[T-\mathbb{E}[T]\right]\left[\mathbb{E}[T]-\theta\right] + \mathbb{E}\left[(\mathbb{E}[T]-\theta)^2\right] \\
&= \mathbb{V}\mathrm{ar}[T] + \left[\mathbb{E}[T]-\theta\right]^2 \\
&= \text{Variance de l'estimateur + carré de son biais}
\end{aligned}$$

Si T est un estimateur sans biais, $EQM[T] = \mathbb{V}\mathrm{ar}[T]$. Pour minimiser l'EQM, on a donc intérêt à ce qu'un estimateur soit sans biais et de faible variance. Par ailleurs, on en déduit immédiatement que **de deux estimateurs sans biais, le meilleur est celui qui a la plus petite variance**. On voit également que l'EQM met en jeu un *compromis biais-variance*, qui s'avèrera être un point clé de nombreux problèmes statistiques et en apprentissage automatique.

La variance d'un estimateur mesure sa variabilité. Si l'estimateur est sans biais, cette variabilité est autour de θ. Si on veut estimer correctement θ, il ne faut pas que cette variabilité soit trop forte.

En pratique, si on observe plusieurs jeux de données similaires, on obtient une estimation de θ pour chacun d'entre eux. Si l'estimateur est sans biais, leur moyenne sera très proche de θ. Si l'estimateur est de forte variance, ces estimations seront éloignées les unes des autres, ce qui fait que chacune de ces estimations peut potentiellement être loin de θ. Si l'estimateur est de faible variance, ces estimations seront toutes proches les unes des autres, mais elles peuvent toutes être éloignées de θ. Si l'estimateur est à la fois sans biais et de faible variance, les estimations de θ seront proches les unes des autres et proches de θ, ce qui fait que n'importe laquelle d'entre elle sera une bonne estimation de θ. C'est évidemment cette dernière situation qui est la plus favorable.

Plaçons nous maintenant dans un modèle d'échantillon de taille n et notons T_n un estimateur. Il est logique de s'attendre à ce que, plus la taille des données augmente, plus on a d'information sur le phénomène aléatoire observé, donc meilleure sera l'estimation. En théorie, avec une observation infinie, on devrait pouvoir estimer θ sans aucune erreur. On peut traduire cette affirmation par le fait que le risque de l'estimateur T_n doit tendre vers 0 quand la taille n de l'échantillon tend vers l'infini. Cela revient à dire que l'estimateur T_n doit converger, en un certain sens, vers θ.

Il s'agit en fait d'étudier la convergence de la suite de variables aléatoires $\{T_n\}_{n\geq 1}$ vers la constante θ. On a vu dans le chapitre 4 qu'il existait plusieurs types de convergence de suites de variables aléatoires. Il existe donc plusieurs types de convergence d'estimateurs.

Définition 5.5.4. L'estimateur T_n de θ est dit :

1. **convergent en probabilité** ou **consistant** ssi $\{T_n\}_{n\geq 1}$ converge en probabilité vers θ :

$$T_n \xrightarrow[n\to\infty]{\mathbb{P}} \theta \text{ i.e. } \forall \epsilon > 0,\ \lim_{n\to\infty} \mathbb{P}\left(|T_n - \theta| > \epsilon\right) = 0$$

2. **convergent en moyenne quadratique** ssi $\{T_n\}_{n\geq 1}$ converge en moyenne quadratique (en moyenne L^2) vers θ :

$$T_n \xrightarrow[n\to\infty]{L^2} \theta \text{ i.e. } \lim_{n\to\infty} EQM[T_n] = \lim_{n\to\infty} \mathbb{E}\left[(T_n-\theta)^2\right] = \lim_{n\to\infty}\left[\mathbb{V}\mathrm{ar}[T_n] + \left[\mathbb{E}[T_n]-\theta\right]^2\right] = 0$$

3. **convergent presque sûrement** ou **fortement consistant** ssi $\{T_n\}_{n\geq 1}$ converge presque sûrement vers θ :

$$T_n \xrightarrow[n\to\infty]{p.s.} \theta \text{ i.e. } \mathbb{P}\left(\left\{\omega\,;\ \lim_{n\to\infty} X_n(\omega) = \theta\right\}\right) = 1$$

En général, on s'intéresse prioritairement à la convergence en moyenne quadratique. Alors, si l'estimateur T_n est sans biais, il sera convergent en moyenne quadratique si et seulement si sa variance tend vers 0 quand n tend vers l'infini.

Finalement, on considèrera que le meilleur estimateur possible de θ est un **estimateur sans biais et de variance minimale (ESBVM)**. Le chapitre 6 explique comment le calculer, s'il existe. Dans ce chapitre, on verra également la notion d'**efficacité d'un estimateur**, qui est très liée à celle d'ESBVM.

Remarque 5.5.1. Ce n'est pas parce que T_n est un bon estimateur de θ que $\varphi(T_n)$ est un bon estimateur de $\varphi(\theta)$. Par exemple, il est fréquent d'avoir $\mathbb{E}[T_n] = \theta$ et $\mathbb{E}\left[\varphi(T_n)\right] \neq \varphi(\theta)$.

En PS1, on a étudié la qualité de l'estimation de l'espérance et de la variance d'une loi de probabilité par respectivement la moyenne et la variance empirique. Il s'agit donc de cas particuliers d'estimation par la méthode des moments. Revenons sur ces résultats.

On se place dans un modèle paramétrique d'échantillon $(\mathcal{X}, \mathcal{A}, \{\mathbb{P}_\theta\,;\,\theta \in \Theta \subset \mathbb{R}\})^n$. Si $\theta = \mathbb{E}[X_1]$, alors l'EMM de θ est $\tilde{\theta}_n = \bar{X}_n$. On a :

$$\mathbb{E}[\bar{X}_n] = \mathbb{E}\big[\frac{1}{n}\sum_{i=1}^n X_i\big] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = \frac{1}{n}\,n\theta = \theta$$

Donc $\bar{X}_n$ est un estimateur sans biais de $\theta = \mathbb{E}[X_1]$. Sa variance est :

$$\mathbb{V}\mathrm{ar}[\bar{X}_n] = \mathbb{V}\mathrm{ar}\big[\frac{1}{n}\sum_{i=1}^n X_i\big] = \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}\mathrm{ar}[X_i] = \frac{\mathbb{V}\mathrm{ar}[X_1]}{n}$$

car les X_i sont indépendantes, donc la variance de leur somme est égale à la somme de leurs variances, qui sont toutes égales à $\mathbb{V}\mathrm{ar}[X_1]$. $\mathbb{V}\mathrm{ar}[\bar{X}_n]$ tend vers 0 quand n tend vers l'infini. Par conséquent $\bar{X}_n$est un estimateur convergent en moyenne quadratique de $\mathbb{E}[X_1]$. Par ailleurs, la loi forte des grands nombres dit que $\bar{X}_n$ converge presque sûrement vers $\mathbb{E}[X_1]$. Par conséquent :

Proposition 5.5.1. *La moyenne empirique $\bar{X}_n$ est un estimateur de $\mathbb{E}[X_1]$ sans biais, convergent en moyenne quadratique et presque sûrement.*

On considère maintenant l'estimation de la variance de la loi des X_i par la variance empirique de l'échantillon $S_n^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2$. Déterminons le biais de cet estimateur.

$$\begin{aligned}
\mathbb{E}[S_n^2] &= \mathbb{E}\big[\frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2\big] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i^2] - \mathbb{E}[\bar{X}_n^2] = \mathbb{E}[X_1^2] - \mathbb{E}[\bar{X}_n^2] \\
&= \mathbb{V}\mathrm{ar}[X_1] + \mathbb{E}[X_1]^2 - \mathbb{V}\mathrm{ar}[\bar{X}_n] - \mathbb{E}[\bar{X}_n]^2 \\
&= \mathbb{V}\mathrm{ar}[X_1] + \mathbb{E}[X_1]^2 - \frac{\mathbb{V}\mathrm{ar}[X_1]}{n} - \mathbb{E}[X_1]^2 = \big(1 - \frac{1}{n}\big)\mathbb{V}\mathrm{ar}[X_1] \\
&= \frac{n-1}{n}\,\mathbb{V}\mathrm{ar}[X_1] \neq \mathbb{V}\mathrm{ar}[X_1]
\end{aligned}$$

Donc, contrairement à ce qu'on pourrait croire, la variance empirique S_n^2 n'est pas un estimateur sans biais de $\mathbb{V}\mathrm{ar}[X_1]$. Cet estimateur n'est qu'asymptotiquement sans biais.

En revanche, on voit que $\mathbb{E}\left[\frac{n}{n-1}\,S_n^2\right] = \frac{n}{n-1}\mathbb{E}[S_n^2] = \mathbb{V}\mathrm{ar}[X_1]$. On pose donc ${S'}_n^2 = \frac{n}{n-1}\,S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. ${S'}_n^2$ est appelée **variance estimée** de l'échantillon. Le résultat précédent montre que c'est un estimateur sans biais de $\mathbb{V}\mathrm{ar}[X_1]$. Par ailleurs, on montre que

$$\mathbb{V}\mathrm{ar}[{S'}_n^2] = \frac{1}{n(n-1)}\left[(n-1)\mathbb{E}\left[(X_1 - \mathbb{E}[X_1])^4\right] - (n-3)\mathbb{V}\mathrm{ar}[X_1]^2\right]$$

qui tend vers 0 quand n tend vers l'infini. La loi forte des grands nombres permet également de montrer que ${S'}_n^2$ et S_n^2 convergent toutes les deux presque sûrement vers $\mathbb{V}\mathrm{ar}[X_1]$. Par conséquent :

Proposition 5.5.2. *La* **variance estimée** ${S'}_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ *est un estimateur de* $\mathbb{V}\mathrm{ar}[X_1]$ *sans biais, convergent en moyenne quadratique et presque sûrement.*

C'est pour cela que la commande `var(x)` en `R` donne la variance estimée, et non pas la variance empirique de l'échantillon `x`.

Remarque 5.5.2. Ces résultats sont non paramétriques : ils sont valables quelle que soit la loi des X_i dans L^1.

Remarque 5.5.3. On n'a pas de résultat général sur la qualité de S_n comme estimateur de l'écart-type de la loi, $\sigma[X_1] = \sqrt{\mathbb{V}\mathrm{ar}[X_1]}$. A priori, ni S_n ni S'_n ne sont des estimateurs sans biais de $\sigma[X_1]$.

Remarque 5.5.4. Le simple exemple de la variance montre qu'un estimateur des moments n'est pas forcément sans biais. On peut montrer que, sous des hypothèses légères vérifiées par la majeure partie des lois de probabilité, un EMM est asymptotiquement sans biais et convergent presque sûrement.

Les propriétés des estimateurs de maximum de vraisemblance seront étudiées dans le chapitre 7. On montrera que ces estimateurs sont asymptotiquement optimaux.

Chapitre 6

Estimation paramétrique optimale

D'après le chapitre précédent, quand on veut estimer un paramètre θ réel dans un modèle paramétrique, l'idéal est d'avoir un estimateur sans biais et de variance minimale (ESBVM). Dans ce chapitre, nous allons expliquer comment trouver un tel estimateur. Dans un premier temps, cela peut parfois se faire via la notion d'estimateur efficace, qui nécessite d'utiliser l'information de Fisher. Dans un second temps, nous donnerons une procédure générale permettant de trouver un ESBVM, qui nécessite d'utiliser les notions d'exhaustivité et de complétude. Les résultats seront d'abord obtenus pour un paramètre θ réel, puis généralisés pour θ de dimension quelconque.

6.1 Information de Fisher pour un paramètre de dimension 1

Dans un modèle statistique paramétrique $(\mathcal{X}, \mathcal{A}, \{\mathbb{P}_\theta^X\,;\, \theta \in \Theta \ \subset \mathbb{R}^d\})$, l'**information de Fisher** mesure la quantité d'information que les données observées x apportent sur le paramètre inconnu θ. Plus l'information de Fisher est grande, plus on pourra estimer θ avec précision. Quand θ est de dimension $d = 1$, l'information de Fisher est un réel et on l'appelle souvent quantité d'information de Fisher. Quand θ est de dimension $d > 1$, l'information de Fisher est une matrice. Dans cette section, on ne considère que le cas $d = 1$. Pour $d > 1$, une définition plus générale sera donnée dans la section suivante. L'élément aléatoire observé est X et la fonction de vraisemblance est $\mathcal{L}(\theta; X)$.

La quantité d'information de Fisher n'est définie que sous certaines conditions de régularité, qui seront détaillées dans un cadre plus général dans la section suivante (hypothèses 6.2.1). Ces conditions sont vérifiées par la plupart des lois de probabilité usuelles. On les supposera vérifiées dans toute cette section.

Définition 6.1.1. On appelle **score** la variable aléatoire :

$$Z(\theta; X) = \frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta; X)$$

Définition 6.1.2. On appelle **quantité d'information** (de Fisher) sur θ apportée par l'observation X, la quantité :

$$\mathcal{I}(\theta) = \mathbb{V}\mathrm{ar}\left[Z(\theta; X)\right] = \mathbb{V}\mathrm{ar}\left[\frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta; X)\right]$$

Remarque 6.1.1. La dérivation de la log-vraisemblance par rapport à θ sert également pour déterminer un estimateur de maximum de vraisemblance de θ.

Le score est donc une variable aléatoire et la quantité d'information est sa variance. On peut montrer que le score est centré ($\mathbb{E}\left[Z(\theta; X)\right] = 0$) et que l'on a également :

$$\mathcal{I}(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \ln \mathcal{L}(\theta; X)\right]$$

L'intérêt de la quantité d'information de Fisher est qu'elle fournit une borne inférieure pour la variance de n'importe quel estimateur sans biais de θ. Ce résultat s'exprime sous la forme de la propriété suivante :

Proposition 6.1.1. Inégalité de Fréchet-Darmois-Cramer-Rao (FDCR) *: Si la loi des observations vérifie les conditions de régularité, alors pour tout estimateur $T = t(X)$ de θ, on a :*

$$\mathbb{V}\mathrm{ar}[T] \geq \frac{\left[\dfrac{\partial}{\partial\theta}\mathbb{E}[T]\right]^2}{\mathcal{I}(\theta)}$$

Démonstration. Pour simplifier, la démonstration est présentée dans le cadre d'un modèle continu. Dans ce cas, d'après le théorème de transfert **??**, la vraisemblance vérifie que, pour toute fonction φ mesurable, on a :

$$\mathbb{E}\left[\varphi(X)\right] = \int_{\mathcal{X}} \varphi(x)\,\mathcal{L}(\theta; x)\,dx. \tag{6.1.1}$$

Le principe de la démonstration est de calculer la covariance entre l'estimateur et le score :

$$\begin{aligned}
\mathbb{C}\mathrm{ov}\big(T, Z(\theta; X)\big) &= \mathbb{E}\big[TZ(\theta; X)\big] - \mathbb{E}[T]\,\mathbb{E}\big[Z(\theta; X)\big] \\
&= \mathbb{E}\big[TZ(\theta; X)\big] \text{ car le score est centré} \\
&= \mathbb{E}\left[t(X)\,\frac{\partial}{\partial\theta}\ln\mathcal{L}(\theta; X)\right] = \int_{\mathcal{X}} t(x)\,\frac{\partial}{\partial\theta}\ln\mathcal{L}(\theta; x)\,\mathcal{L}(\theta; x)\,dx \\
&= \int_{\mathcal{X}} t(x)\,\frac{\partial}{\partial\theta}\mathcal{L}(\theta; x)\,dx = \frac{\partial}{\partial\theta}\int_{\mathcal{X}} t(x)\,\mathcal{L}(\theta; x)\,dx \\
&= \frac{\partial}{\partial\theta}\,\mathbb{E}[t(X)] = \frac{\partial}{\partial\theta}\,\mathbb{E}[T].
\end{aligned}$$

Dans ce calcul, la possibilité de dérivation sous le signe intégrale utilisée à l'avant-dernière égalité fait partie des hypothèses de régularité.

L'**inégalité de Cauchy-Schwarz** est un cas particulier de l'inégalité de Hölder vue au théorème 2.4.1, avec $p = q = 2$. Elle s'écrit ici :

$$\mathbb{C}\mathrm{ov}\big(T, Z(\theta; X)\big)^2 \leq \mathbb{V}\mathrm{ar}[T]\,\mathbb{V}\mathrm{ar}\big[Z(\theta; X)\big]$$

D'où :

$$\mathbb{V}\mathrm{ar}[T] \geq \frac{\mathbb{C}\mathrm{ov}\big(T, Z(\theta; X)\big)^2}{\mathbb{V}\mathrm{ar}\big[Z(\theta; X)\big]} = \frac{\left[\dfrac{\partial}{\partial\theta}\mathbb{E}[T]\right]^2}{\mathcal{I}(\theta)}.$$

□

L'inégalité FDCR est particulièrement intéressante pour les estimateurs sans biais. En effet, si T est un estimateur sans biais de θ, alors $\mathbb{E}[T] = \theta$, donc $\mathbb{V}\mathrm{ar}[T] \geq \dfrac{1}{\mathcal{I}(\theta)}$. La quantité $\dfrac{1}{\mathcal{I}(\theta)}$ est appelée la **borne de Cramer-Rao**. L'inégalité FDCR dit donc que la variance d'un estimateur sans biais quelconque de θ est forcément supérieure à cette borne.

Définition 6.1.3. On appelle **efficacité** d'un estimateur T la quantité :

$$\mathbb{E}\mathrm{ff}[T] = \frac{\left[\dfrac{\partial}{\partial\theta}\mathbb{E}[T]\right]^2}{\mathcal{I}(\theta)\mathbb{V}\mathrm{ar}[T]}$$

L'inégalité FDCR implique que $0 \leq \mathbb{E}\text{ff}[T] \leq 1$. T est dit un estimateur **efficace** si et seulement si $\mathbb{E}\text{ff}[T] = 1$. On a donc les propriétés suivantes :

- Si T est un estimateur sans biais de θ, $\mathbb{E}\text{ff}[T] = \dfrac{1}{\mathcal{I}(\theta)Var[T]}$.
- Si un estimateur sans biais est efficace, sa variance est égale à la borne de Cramer-Rao, donc c'est forcément un ESBVM.
- Il est possible qu'il n'existe pas d'estimateur sans biais efficace de θ. Alors, s'il existe un ESBVM de θ, sa variance est strictement supérieure à la borne de Cramer-Rao.
- Si la valeur de la borne de Cramer-Rao est très grande, il est impossible d'estimer correctement θ car tous les estimateurs sans biais possibles auront une forte variance.

La façon la plus simple de trouver un ESBVM est donc de trouver un estimateur sans biais efficace. Quand on n'en trouve pas ou quand il n'en existe pas, on obtiendra un ESBVM en utilisant le théorème de Lehmann-Scheffé, présenté en section 6.7.

La définition 6.1.2 de la quantité d'information est une définition générale, applicable quelle que soit la nature de l'observation X. Pour un modèle d'échantillon de taille n ($X = (X_1, \ldots, X_n)$ où les X_i sont sont indépendantes et de même loi), on note $\mathcal{I}_n(\theta)$ la quantité d'information. Il est alors facile de voir que $\mathcal{I}_n(\theta) = n\mathcal{I}_1(\theta)$. Par exemple, pour des variables aléatoires continues de densité f :

$$\begin{aligned}
\mathcal{I}_n(\theta) &= Var\left[\frac{\partial}{\partial\theta}\ln \mathcal{L}(\theta; X_1, \ldots, X_n)\right] = Var\left[\frac{\partial}{\partial\theta}\ln \prod_{i=1}^{n} f(X_i;\theta)\right] \\
&= Var\left[\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\ln f(X_i;\theta)\right] = Var\left[\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f(X_i;\theta)\right] \\
&= \sum_{i=1}^{n} Var\left[\frac{\partial}{\partial\theta}\ln f(X_i;\theta)\right] = n\mathcal{I}_1(\theta)
\end{aligned}$$

Cette propriété traduit l'idée naturelle que, dans un échantillon, chaque observation porte la même quantité d'information sur θ, et que la quantité d'information est additive.

Dans ce modèle d'échantillon, on note T_n un estimateur de θ. Comme dit plus haut, si T_n est sans biais et efficace, c'est un ESBVM de θ. Si on n'a pas d'estimateur sans biais et efficace, on se contente souvent d'un estimateur sans biais et **asymptotiquement efficace**, c'est-à-dire tel que $\lim_{n\to+\infty} \mathbb{E}\text{ff}(T_n) = 1$. On verra dans le chapitre suivant que l'estimateur de maximum de vraisemblance est asymptotiquement sans biais et efficace, ce qui entraine qu'il est asymptotiquement optimal.

Exemple 6.1.1. *Contrôle de qualité.* $X = (X_1, \ldots, X_n)$ où les X_i sont indépendantes et de même loi de Bernoulli $\mathcal{B}(p)$. L'EMM et EMV de p est $\hat{p}_n = \bar{X}_n$. On sait que $\bar{X}_n$ est un estimateur sans biais de $\mathbb{E}[X_i]$. Or l'espérance de la loi $\mathcal{B}(p)$ est p, donc $\hat{p}_n$ est un estimateur sans biais de p.

On sait aussi que $\mathbb{V}\text{ar}[\bar{X}_n] = \dfrac{\mathbb{V}\text{ar}[X]}{n} = \dfrac{p(1-p)}{n}$, donc $\hat{p}_n$ est convergent en moyenne quadratique.
La quantité d'information est :

$$\begin{aligned}
\mathcal{I}_n(p) &= \mathbb{V}\text{ar}\left[\frac{\partial}{\partial p}\ln \mathcal{L}(p; X_1, \ldots, X_n)\right] = \mathbb{V}\text{ar}\left[\frac{\sum_{i=1}^{n} X_i - np}{p(1-p)}\right] = \frac{\mathbb{V}\text{ar}\left[\sum_{i=1}^{n} X_i\right]}{p^2(1-p)^2} \\
&= \frac{np(1-p)}{p^2(1-p)^2} \text{ car } \sum_{i=1}^{n} X_i \text{ est de loi binomiale } \mathcal{B}(n,p) \\
&= \frac{n}{p(1-p)}
\end{aligned}$$

On a donc $\mathbb{V}\text{ar}[\hat{p}_n] = \dfrac{1}{\mathcal{I}_n(p)}$, ce qui prouve que $\hat{p}_n$ est un estimateur efficace. Par conséquent, $\hat{p}_n$ est un ESBVM de p. Autrement dit, la meilleure façon possible d'estimer la probabilité qu'une pièce soit défectueuse est de prendre le pourcentage de pièces défectueuses dans le lot observé. Cela conforte l'intuition.

6.2 Information de Fisher pour un paramètre de dimension quelconque

Dans cette section, nous allons généraliser les résultats de la section précédente au cas où θ est de dimension d quelconque en se plaçant dans un modèle paramétrique $(\mathcal{X}, \mathcal{A}, \{\mathbb{P}_\theta^X \, ; \, \theta \in \Theta \subset \mathbb{R}^d\})$.

Dans ce qui suit, pour simplifier, les calculs seront effectués pour un modèle continu. Mais ils sont généralisables à n'importe quel modèle statistique paramétrique, en utilisant la définition 5.4.3 de la vraisemblance.

La quantité d'information n'est définie que sous les conditions de régularité suivantes.

Hypothèses 6.2.1. Conditions de régularité pour la définition de l'information de Fisher.

- Le support de $\mathbb{P}_\theta^X$ ne dépend pas de θ (ce qui, par exemple, exclut la loi uniforme sur $[0, \theta]$).
- $\forall \theta, \forall x, \mathcal{L}(\theta; x) > 0$.
- $\ln \mathcal{L}(\theta; x)$ est dérivable 2 fois par rapport à θ.
- On peut dériver 2 fois sous le signe somme par rapport à θ : pour toute fonction mesurable g,

$$\frac{\partial}{\partial \theta} \int_A g(x)\, \mathcal{L}(\theta; x)\, dx = \int_A g(x)\, \frac{\partial}{\partial \theta} \mathcal{L}(\theta; x)\, dx$$

et

$$\frac{\partial^2}{\partial \theta^2} \int_A g(x)\, \mathcal{L}(\theta; x)\, dx = \int_A g(x)\, \frac{\partial^2}{\partial \theta^2} \mathcal{L}(\theta; x)\, dx.$$

Dans la suite, on suppose que les hypothèses 6.2.1 sont vérifiées. On peut alors définir les quantités suivantes.

6.2.1 Score et matrice d'information

Dans les calculs algébriques, les vecteurs sont des vecteurs colonnes. Le paramètre θ s'écrit donc $\theta = (\theta_1, \ldots, \theta_d)^T$. Quand on estime un paramètre θ de dimension d, les notions usuelles liées à l'estimation s'écrivent sous forme vectorielle. Par exemple, un estimateur T de θ sera un vecteur aléatoire de dimension d.

Le vecteur aléatoire $T = (T_1, \ldots, T_d)^T$ est un estimateur sans biais de θ si $\mathbb{E}[T] = \theta$, ce qui s'écrit vectoriellement $(\mathbb{E}[T_1], \ldots, \mathbb{E}[T_d])^T = (\theta_1, \ldots, \theta_d)^T$ ou $\forall j \in \{1, \ldots, d\}, \mathbb{E}[T_j] = \theta_j$.

Pour θ de dimension d, les notions de score et de quantité d'information de Fisher se généralisent de la manière suivante.

Définition 6.2.1. Le **score** est le gradient de la log-vraisemblance :

$$Z(\theta; X) = \nabla_\theta \ln \mathcal{L}(\theta; X) = (Z_1(\theta; X), \ldots, Z_d(\theta; X))^T$$

où $\forall j \in \{1, \ldots, d\},\ Z_j(\theta; X) = \dfrac{\partial}{\partial \theta_j} \ln \mathcal{L}(\theta; X)$.

Le score est un vecteur aléatoire de dimension d. L'estimateur de maximum de vraisemblance $\hat{\theta}$ de θ est la valeur de θ qui annule le score : $Z(\hat{\theta}; X) = 0$.

Définition 6.2.2. La **matrice d'information de Fisher** $\mathcal{I}(\theta)$ est la matrice de covariance du score, de terme général

$$\forall (j,k) \in \{1,\ldots,d\}^2,\ \mathcal{I}_{jk}(\theta) = \mathbb{C}\mathrm{ov}\big[Z_j(\theta;X); Z_k(\theta;X)\big].$$

Quand $\theta \in \mathbb{R}$, on retrouve bien $Z(\theta;X) = \dfrac{\partial}{\partial\theta}\ln\mathcal{L}(\theta;X)$ et $\mathcal{I}(\theta) = \mathbb{V}\mathrm{ar}[Z(\theta;X)] = \mathbb{V}\mathrm{ar}\left[\dfrac{\partial}{\partial\theta}\ln\mathcal{L}(\theta;X)\right]$.

Les propriétés vues dans la section 6.1 pour $d = 1$ se généralisent au cas $d > 1$.

Proposition 6.2.1.

1. *Le score est centré :* $\mathbb{E}[Z(\theta;X)] = 0$.
2. $\forall (j,k) \in \{1,\ldots,d\}^2$, $\mathcal{I}_{jk}(\theta) = -\mathbb{E}\left[\dfrac{\partial^2}{\partial\theta_j\partial\theta_k}\ln\mathcal{L}(\theta;X)\right]$
3. *Pour les modèles d'échantillon de taille n, la matrice d'information est $\mathcal{I}_n(\theta) = n\mathcal{I}_1(\theta)$.*

6.2.2 Borne de Cramer-Rao et efficacité

L'inégalité FDCR vue dans la proposition 6.1.1 pour $\theta \in \mathbb{R}$ se généralise pour θ de dimension quelconque, mais la formulation est nettement plus complexe.

Théorème 6.2.1. Inégalité de Fréchet-Darmois-Cramer-Rao (FDCR). *On considère un modèle paramétrique $(\mathcal{X}, \mathcal{A}, \{\mathbb{P}_\theta^X\,;\,\theta \in \Theta \subset \mathbb{R}^d\})$ vérifiant les hypothèses de cette section et tel que la matrice d'information $\mathcal{I}(\theta)$ soit inversible.*

Soit t une statistique à valeurs dans $\mathbb{R}^q$, K_T la matrice de covariance de T et Δ la matrice de terme général $\Delta_{ij} = \dfrac{\partial}{\partial\theta_j}\mathbb{E}[T_i]$, $1 \leq i \leq q$, $1 \leq j \leq d$.

Alors $\forall\theta \in \mathbb{R}^d$, la matrice $K_T - \Delta\,\mathcal{I}^{-1}(\theta)\,\Delta^T$ est semi-définie positive.

Rappel : La matrice M est semi-définie positive si et seulement si $\forall x \neq 0, x^T M x \geq 0$.

Quand $d = q = 1$, $K_T = \mathbb{V}\mathrm{ar}[T]$ et $\Delta = \dfrac{\partial}{\partial\theta}\mathbb{E}[T] \in \mathbb{R}$. Alors on obtient :

$$\mathbb{V}\mathrm{ar}[T] - \frac{\left[\dfrac{\partial}{\partial\theta}\mathbb{E}[T]\right]^2}{\mathcal{I}(\theta)} \geq 0.$$

C'est bien le résultat attendu.

Quand $\theta \in \mathbb{R}^d$, l'inégalité FDCR appliquée aux termes diagonaux de K_T permet d'obtenir une borne inférieure pour la variance de chaque composante de T :

Proposition 6.2.2. $\forall i \in \{1,\ldots,q\}$, *on a :*

$$\mathbb{V}\mathrm{ar}[T_i] \geq \sum_{j=1}^{d}\sum_{k=1}^{d}\mathcal{I}^{-1}(\theta)_{jk}\,\frac{\partial\mathbb{E}[T_i]}{\partial\theta_j}\,\frac{\partial\mathbb{E}[T_i]}{\partial\theta_k}.$$

En particulier, si T est un estimateur sans biais de θ, on a pour tout i, $\mathbb{E}[T_i] = \theta_i$. Donc $\dfrac{\partial\mathbb{E}[T_i]}{\partial\theta_j} = \delta_{ij} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{sinon} \end{cases}$ $(\Delta = Id)$, d'où $\mathbb{V}\mathrm{ar}[T_i] \geq \mathcal{I}^{-1}(\theta)_{ii}$, qui est la borne de Cramer-Rao.

L'estimateur T est efficace si l'inégalité FDCR est une égalité.

Définition 6.2.3. Un estimateur sans biais T est **efficace** si et seulement si $K_T = \mathcal{I}^{-1}(\theta)$. Alors, pour tout i, $\mathbb{V}\mathrm{ar}[T_i] = \mathcal{I}^{-1}(\theta)_{ii}$.

Les notions d'information de Fisher et d'efficacité d'estimateur permettent dans certains cas d'obtenir un ESBVM d'un paramètre, mais malheureusement, cela n'est pas le cas pour tous les modèles statistiques. Pour aller plus loin, il va falloir utiliser les notions développées dans les sections suivantes.

6.3 Exhaustivité

On considère un modèle statistique paramétrique $(\mathcal{X}, \mathcal{A}, \{\mathbb{P}_\theta^X\,;\, \theta \in \Theta \subset \mathbb{R}^d\})$. On cherche à obtenir le maximum de connaissance possible sur le paramètre θ à partir de l'observation $x \in \mathcal{X}$. Souvent, x est un vecteur $(x_1, \ldots, x_n)$ et n est très grand. Il est alors intéressant de réduire les données en les résumant par une statistique $t(x)$ de dimension très inférieure à n (par exemple, la moyenne empirique $\bar{x}_n$ est un résumé de x de dimension 1). Il est logique de s'attendre à ce que le résumé $t(x)$ des observations contienne moins d'information sur θ que l'ensemble des données initiales. Or il existe des statistiques qui résument les observations tout en conservant l'intégralité de l'information sur θ, les statistiques exhaustives.

Définition 6.3.1. Une statistique t est **exhaustive** pour θ si et seulement si la loi de probabilité conditionnelle de X sachant $[t(X) = t_0]$ ne dépend pas de θ.

Justification. Si la loi de X sachant $[t(X) = t_0]$ ne dépend pas de θ, cela signifie que, quand on connait le résumé de l'observation $t(x)$, la connaissance de la totalité de l'observation x n'apporte aucun renseignement supplémentaire sur θ. Donc la totalité de l'information sur θ est contenue dans $t(x)$. Par conséquent, il faut s'attendre à ne se servir que de $t(x)$ (au lieu de x tout entier) pour estimer θ.

Exemple 6.3.1. *Contrôle de qualité*. Le modèle est $\left(\{0,1\}, \mathcal{P}\left(\{0,1\}\right), \{\mathcal{B}(p); p \in [0,1]\}\right)^n$. $x = (x_1, \ldots, x_n)$, où

$$x_i = \begin{cases} 1 \text{ si la } i^{\text{ème}} \text{ pièce est défectueuse} \\ 0 \text{ sinon} \end{cases}$$

Les X_i sont des variables aléatoires indépendantes et de même loi $\mathcal{B}(p)$, où p est la probabilité qu'une pièce soit défectueuse. Il semble évident que, pour avoir toute l'information sur p, il est inutile de savoir, pour chaque pièce contrôlée, si elle est défectueuse ou pas. Il suffit de connaître le pourcentage (ou le nombre total) de pièces défectueuses. On doit donc s'attendre à ce que ces quantités soient des statistiques exhaustives. Pour des raisons de simplicité d'écriture, on va montrer que le nombre total de pièces défectueuses $t(x) = \sum_{i=1}^n x_i$ est une statistique exhaustive. Pour cela, il faut montrer que $\mathbb{P}(X = x | t(X) = t_0; p)$ ne dépend pas de p.

On sait que $T = t(X) = \sum_{i=1}^n X_i$ est de loi binomiale $\mathcal{B}(n,p)$. Alors :

$$\mathbb{P}(X = x | T = t_0; p) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n | \sum_{i=1}^n X_i = t_0; p)$$

$$= \frac{\mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n, \sum_{i=1}^n X_i = t_0; p\right)}{\mathbb{P}\left(\sum_{i=1}^n X_i = t_0; p\right)} = \begin{cases} 0 & \text{si } \sum_{i=1}^n x_i \neq t_0 \\ \dfrac{\mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n; p\right)}{\mathbb{P}\left(\sum_{i=1}^n X_i = t_0; p\right)} & \text{si } \sum_{i=1}^n x_i = t_0 \end{cases}$$

Or $\mathbb{P}(X_i = x_i; p) = \begin{cases} p & \text{si } x_i = 1 \\ 1-p & \text{si } x_i = 0 \end{cases} = p^{x_i}(1-p)^{1-x_i}$. Et comme les X_i sont indépendantes, on a :

$$
\begin{aligned}
\frac{\mathbb{P}\left(X_1=x_1,\ldots,X_n=x_n;p\right)}{\mathbb{P}\left(\sum\limits_{i=1}^{n} X_i=t_0;p\right)} &= \frac{\prod\limits_{i=1}^{n}\mathbb{P}\left(X_i=x_i;p\right)}{\mathbb{P}\left(T=t_0;p\right)} = \frac{\prod\limits_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}}{\binom{n}{t_0}\, p^{t_0}\,(1-p)^{n-t_0}} \\
&= \frac{p^{\sum\limits_{i=1}^{n} x_i}\,(1-p)^{n-\sum\limits_{i=1}^{n} x_i}}{\binom{n}{t_0}\, p^{t_0}\,(1-p)^{n-t_0}} = \frac{1}{\binom{n}{t_0}} \text{ si } \sum_{i=1}^{n} x_i = t_0.
\end{aligned}
$$

Donc $\mathbb{P}(X=x|T=t_0;p)=\left\{\begin{array}{cl} 0 & \text{si } \sum\limits_{i=1}^{n} x_i \neq t_0 \\ \dfrac{1}{\binom{n}{t_0}} & \text{si } \sum\limits_{i=1}^{n} x_i = t_0 \end{array}\right. .$

On reconnait la loi uniforme sur $\left\{(x_1,\ldots,x_n)\in\{0,1\}^n;\ \sum\limits_{i=1}^{n} x_i=t_0\right\}$.

La loi conditionnelle de X sachant $[T=t_0]$ ne dépend pas de p, donc $t(x)=\sum\limits_{i=1}^{n} x_i$ est une statistique exhaustive pour p.

La vérification de la propriété définissant les statistiques exhaustives nécessitant d'avoir une idée préalable de la forme de cette statistique, il est plus pratique d'utiliser le théorème de Fisher-Neyman, qui caractérise très simplement l'exhaustivité.

Théorème 6.3.1. Théorème de factorisation de Fisher-Neyman. *Pour qu'une statistique t soit exhaustive pour θ, il faut et il suffit qu'il existe deux fonctions mesurables g et h telles que :*

$$\forall\, x \in \mathcal{X},\ \forall\, \theta \in \Theta,\ \ \mathcal{L}(\theta;x) = g\left(t(x);\theta\right)\ h(x).$$

Autrement dit, ce qui dépend de θ et de x ne dépend de x qu'à travers $t(x)$.

Démonstration. Effectuons la démonstration dans le cas d'un modèle discret. On a donc $\mathcal{L}(\theta;x)=\mathbb{P}(X=x;\theta)$. Notons $T=t(X)$.

($\Rightarrow$) Si t est exhaustive, $\mathbb{P}(X=x|T=t_0;\theta)$ ne dépend pas de θ. Par conséquent :

$$
\begin{aligned}
\mathcal{L}(\theta;x) &= \mathbb{P}(X=x;\theta) = \mathbb{P}\left(\{X=x\}\cap\{t(X)=t(x)\};\theta\right) \\
&= \mathbb{P}\left(\{X=x\}\cap\{T=t(x)\};\theta\right) = \mathbb{P}(X=x|T=t(x);\theta)\,\mathbb{P}(T=t(x);\theta) \\
&= \mathbb{P}(X=x|T=t(x))\,\mathbb{P}(T=t(x);\theta) = h(x)\,\mathbb{P}(T=t(x);\theta)
\end{aligned}
$$

qui est bien de la forme $g\left(t(x);\theta\right)\ h(x)$.

($\Leftarrow$) On suppose que $\mathcal{L}(\theta;x)=\mathbb{P}(X=x;\theta)=g\left(t(x);\theta\right)\ h(x)$. Il faut montrer qu'alors $\mathbb{P}(X=x|T=t_0;\theta)$ ne dépend pas de θ. On a :

$$
\begin{aligned}
\mathbb{P}(X=x|T=t_0;\theta) &= \frac{\mathbb{P}\left(\{X=x\}\cap\{T=t_0\};\theta\right)}{\mathbb{P}(T=t_0;\theta)} = \frac{\mathbb{P}\left(\{X=x\}\cap\{t(X)=t_0\};\theta\right)}{\mathbb{P}(T=t_0;\theta)} \\
&= \left\{\begin{array}{cl} 0 & \text{si } t(x)\neq t_0 \\ \dfrac{\mathbb{P}(X=x;\theta)}{\mathbb{P}(T=t_0;\theta)} & \text{si } t(x)=t_0 \end{array}\right.
\end{aligned}
$$

Or $\mathbb{P}(T=t_0;\theta)=\mathbb{P}(t(X)=t_0;\theta)=\sum\limits_{y;t(y)=t_0}\mathbb{P}(X=y;\theta)$.

Donc, pour $t(x) = t_0$, on a :

$$\begin{aligned}
\mathbb{P}(X = x | T = t_0; \theta) &= \frac{\mathbb{P}(X = x; \theta)}{\sum\limits_{y; t(y) = t_0} \mathbb{P}(X = y; \theta)} = \frac{g\,(t(x); \theta)\ h(x)}{\sum\limits_{y; t(y) = t_0} g\,(t(y); \theta)\ h(y)} \\
&= \frac{g\,(t_0; \theta)\ h(x)}{\sum\limits_{y; t(y) = t_0} g\,(t_0; \theta)\ h(y)} = \frac{h(x)}{\sum\limits_{y; t(y) = t_0} h(y)}
\end{aligned}$$

qui ne dépend pas de θ. Donc t est exhaustive, d'où le théorème. □

Exemple 6.3.2. *Contrôle de qualité.* On a vu que :

$$\mathcal{L}(p; x_1, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} = p^{\sum\limits_{i=1}^{n} x_i} (1-p)^{n - \sum\limits_{i=1}^{n} x_i}.$$

C'est de la forme $g(\sum\limits_{i=1}^{n} x_i; p)$, donc on retrouve immédiatement que $\sum\limits_{i=1}^{n} x_i$ est une statistique exhaustive. Cette manière de déterminer une statistique exhaustive est beaucoup plus simple que l'utilisation de la définition faite plus haut. De plus, elle ne nécessite pas de connaître à l'avance la statistique t.

Exemple 6.3.3. *Echantillon de loi normale $\mathcal{N}(m; \sigma^2)$.* On suppose que $X = (X_1, \ldots, X_n)$, où les X_i sont indépendantes et de même loi $\mathcal{N}(m; \sigma^2)$. La vraisemblance est :

$$\begin{aligned}
\mathcal{L}(m, \sigma^2; x_1, \ldots, x_n) &= \prod_{i=1}^{n} f_{X_i}\left(x_i; m, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\dfrac{(x_i - m)^2}{2\sigma^2}} \\
&= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n}\, e^{-\dfrac{1}{2\sigma^2} \sum\limits_{i=1}^{n} (x_i - m)^2} \\
&= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n}\, e^{-\dfrac{1}{2\sigma^2}\left[\sum\limits_{i=1}^{n} x_i^2 - 2m \sum\limits_{i=1}^{n} x_i + nm^2\right]}
\end{aligned}$$

qui est de la forme $g\left((\sum\limits_{i=1}^{n} x_i, \sum\limits_{i=1}^{n} x_i^2); m, \sigma^2\right)$. Donc le couple $\left(\sum\limits_{i=1}^{n} x_i, \sum\limits_{i=1}^{n} x_i^2\right)$ est une statistique exhaustive pour le paramètre $\theta = (m, \sigma^2)$ d'un échantillon de loi normale.

Proposition 6.3.1. *Si t est exhaustive et si $t = \varphi \circ s$, alors s est exhaustive.*

Démonstration. t est exhaustive donc

$$\mathcal{L}(\theta; x) = g\,(t(x); \theta)\ h(x) = g\,(\varphi[s(x)]; \theta)\ h(x) = G\,(s(x); \theta)\ h(x)$$

donc s est exhaustive. □

Exemple 6.3.4. *Echantillon de loi normale.* $\left(\sum\limits_{i=1}^{n} x_i, \sum\limits_{i=1}^{n} x_i^2\right) = \varphi(\bar{x}_n, s_n^2)$, donc $(\bar{x}_n, s_n^2)$ est une statistique exhaustive pour (m, σ^2) (c'est la statistique de maximum de vraisemblance).

Remarque 6.3.1. Si t est exhaustive, $\varphi \circ t$ ne l'est pas forcément ! Par exemple, $\varphi(\bar{x}_n, s_n^2) = \bar{x}_n$ n'est pas exhaustive pour (m, σ^2).

Proposition 6.3.2. *Si t est une statistique exhaustive et si $\hat{\theta}$ est la statistique de maximum de vraisemblance, alors il existe une fonction φ telle que $\hat{\theta} = \varphi \circ t$.*

Démonstration. t est exhaustive donc $\mathcal{L}(\theta; x) = g\left(t(x); \theta\right) h(x)$. h n'intervient pas dans la maximisation de cette fonction par rapport à θ, donc la statistique de maximum de vraisemblance ne dépend de x qu'à travers $t(x)$. Par conséquent, il existe une fonction φ telle que $\hat{\theta} = \varphi \circ t$. $\square$

C'est bien le cas de la loi normale avec $t(x) = \left(\sum\limits_{i=1}^{n} x_i, \sum\limits_{i=1}^{n} x_i^2\right)$ et $\hat{\theta}(x) = (\bar{x}_n, s_n^2)$.

La statistique de maximum de vraisemblance est fonction d'une statistique exhaustive, mais elle n'est pas forcément exhaustive elle-même.

Il existe un lien entre information de Fisher et exhaustivité, donné par la propriété suivante. Pour $\theta \in \mathbb{R}$, soit $\mathcal{I}(\theta)$ la quantité d'information d'un modèle paramétrique $\left(\mathcal{X}, \mathcal{A}, \left\{\mathbb{P}_\theta^X ; \theta \in \Theta\right\}\right)$. Soit t une statistique, application mesurable de $(\mathcal{X}, \mathcal{A})$ dans $(\mathcal{Y}, \mathcal{B})$. Soit $\mathbb{P}_\theta^T$ la loi de probabilité de T. L'observation de $t(x)$ conduit au modèle statistique $\left(\mathcal{Y}, \mathcal{B}, \left\{\mathbb{P}_\theta^T ; \theta \in \Theta\right\}\right)$. Soit $\mathcal{I}_t(\theta)$ la quantité d'information de ce modèle.

Proposition 6.3.3. *Dégradation de l'information : pour toute statistique t, $\mathcal{I}_t(\theta) \leq \mathcal{I}(\theta)$.*
Information et exhaustivité : $\mathcal{I}_t(\theta) = \mathcal{I}(\theta) \Leftrightarrow t$ est exhaustive.

Autrement dit, si on résume les données x par une statistique $t(x)$, on perd de l'information, sauf si la statistique est exhaustive.

En fait, on peut caractériser facilement les lois de probabilité pour lesquelles les modèles d'échantillon admettent une statistique exhaustive : celles qui appartiennent à la famille exponentielle.

6.4 La famille exponentielle

Définition 6.4.1. Soit X une variable aléatoire réelle, dont la loi de probabilité dépend d'un paramètre $\theta \in \mathbb{R}^d$. On dit que la loi de X **appartient à la famille exponentielle** si et seulement si la fonction de vraisemblance s'écrit :

$$\mathcal{L}(\theta; x) = e^{\sum\limits_{j=1}^{d} a_j(x)\alpha_j(\theta) + b(x) + \beta(\theta)}$$

Comme on n'a considéré qu'une seule variable aléatoire réelle, pour un modèle discret $\mathcal{L}(\theta; x) = \mathbb{P}(X = x; \theta)$ et pour un modèle continu, $\mathcal{L}(\theta; x) = f_X(x; \theta)$.

La plupart des lois usuelles appartiennent à la famille exponentielle :

- *Loi de Bernoulli $\mathcal{B}(p)$* :

$$\begin{aligned}\mathbb{P}(X = x; p) &= \left\{\begin{array}{ll} p & \text{si } x = 1 \\ 1-p & \text{si } x = 0 \end{array}\right. = p^x(1-p)^{1-x} = e^{x \ln p + (1-x)\ln(1-p)} \\ &= e^{x[\ln p - \ln(1-p)] + \ln(1-p)} = e^{x \ln \frac{p}{1-p} + \ln(1-p)}\end{aligned}$$

  qui est de la forme souhaitée avec $d = 1$, $a(x) = x$, $\alpha(p) = \ln \dfrac{p}{1-p}$, $b(x) = 0$ et $\beta(p) = \ln(1-p)$.

- *Loi exponentielle* $exp(\lambda)$:
$$f_X(x;\lambda) = \lambda\, e^{-\lambda x} = e^{-\lambda x + \ln \lambda}$$
qui est de la forme souhaitée avec $d = 1$, $a(x) = x$, $\alpha(\lambda) = -\lambda$, $b(x) = 0$ et $\beta(\lambda) = \ln \lambda$.

- *Loi normale* $\mathcal{N}(m, \sigma^2)$:
$$f_X\left(x; m, \sigma^2\right) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\dfrac{(x-m)^2}{2\sigma^2}} = e^{-\dfrac{x^2}{2\sigma^2} + \dfrac{mx}{\sigma^2} - \dfrac{m^2}{2\sigma^2} - \ln \sigma\sqrt{2\pi}}$$
qui est de la forme souhaitée avec $d = 2$, $a_1(x) = x^2$, $\alpha_1(m, \sigma^2) = -\dfrac{1}{2\sigma^2}$, $a_2(x) = x$, $\alpha_2(m, \sigma^2) = \dfrac{m}{\sigma^2}$, $b(x) = 0$ et $\beta(m, \sigma^2) = -\dfrac{m^2}{2\sigma^2} - \ln \sigma\sqrt{2\pi}$.

Mais par exemple, la loi de Weibull $\mathcal{W}(\eta, \beta)$ n'appartient pas à la famille exponentielle :
$$f_X\left(x; \eta, \beta\right) = \beta\, \frac{x^{\beta-1}}{\eta^\beta}\, e^{-\left(\dfrac{x}{\eta}\right)^\beta} = e^{-\dfrac{x^\beta}{\eta^\beta} + (\beta - 1)\ln x - \beta \ln \eta + \ln \beta}$$

Le terme x^β fait que $\dfrac{x^\beta}{\eta^\beta}$ ne peut pas être mis sous la forme $a(x)\alpha(\eta, \beta)$, donc la loi de Weibull n'appartient pas à la famille exponentielle.

Le lien entre famille exponentielle et exhaustivité est donné par le théorème de Darmois :

Théorème 6.4.1. Théorème de Darmois. *Dans un modèle d'échantillon* $(\mathcal{X}, \mathcal{A}, \left\{P_\theta^X\,;\, \theta \in \Theta\, \subset \mathbb{R}^d\right\})^n$, *où le support de la loi des observations ne dépend pas de* θ, *il existe une statistique exhaustive si et seulement si cette loi appartient à la famille exponentielle. Alors* $t(x) = \left(\sum\limits_{i=1}^n a_1(x_i), \ldots, \sum\limits_{i=1}^n a_d(x_i)\right)$ *est une statistique exhaustive.*

Démonstration. On effectue la démonstration pour des lois continues.

($\Leftarrow$) Si la loi des observations appartient à la famille exponentielle, la fonction de vraisemblance est :
$$\begin{aligned}
\mathcal{L}(\theta; x_1, \ldots, x_n) &= \prod_{i=1}^n f_{X_i}(x_i;\theta) = \prod_{i=1}^n e^{\sum\limits_{j=1}^d a_j(x_i)\alpha_j(\theta) + b(x_i) + \beta(\theta)} \\
&= e^{\sum\limits_{i=1}^n \sum\limits_{j=1}^d a_j(x_i)\alpha_j(\theta) + \sum\limits_{i=1}^n b(x_i) + n\beta(\theta)} \\
&= e^{\sum\limits_{j=1}^d \alpha_j(\theta) \sum\limits_{i=1}^n a_j(x_i) + \sum\limits_{i=1}^n b(x_i) + n\beta(\theta)}
\end{aligned}$$

Le théorème de Fisher-Neyman permet alors d'en déduire que $t(x) = \left(\sum\limits_{i=1}^n a_1(x_i), \ldots, \sum\limits_{i=1}^n a_d(x_i)\right)$ est une statistique exhaustive pour θ.

($\Rightarrow$) Montrons la réciproque pour $d = 1$, c'est-à-dire $\theta \in \mathbb{R}$. On suppose qu'il existe une statistique exhaustive t. Alors :
$$\mathcal{L}(\theta; x_1, \ldots, x_n) = \prod_{i=1}^n f(x_i;\theta) = g\left(t(x_1, \ldots, x_n); \theta\right)\, h(x_1, \ldots, x_n)$$

Il faut montrer qu'alors forcément $f(x;\theta)$ est de la forme $e^{a(x)\alpha(\theta) + b(x) + \beta(\theta)}$. On a :
$$\ln \mathcal{L}(\theta; x_1, \ldots, x_n) = \sum_{i=1}^n \ln f(x_i;\theta) = \ln g\left(t(x_1, \ldots, x_n); \theta\right) + \ln h(x_1, \ldots, x_n)$$

Et comme h ne dépend pas de θ :

$$\frac{\partial}{\partial\theta}\ \ln \mathcal{L}(\theta; x_1, \ldots, x_n) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta}\ \ln f(x_i;\theta) = \frac{\partial}{\partial\theta}\ \ln g\left(t(x_1, \ldots, x_n);\theta\right)$$

Pour un i quelconque fixé dans $\{1, \ldots, n\}$, on a :

$$\begin{aligned}\frac{\partial^2}{\partial\theta\partial x_i}\ \ln \mathcal{L}(\theta; x_1, \ldots, x_n) &= \frac{\partial^2}{\partial\theta\partial x_i}\ \ln f(x_i;\theta) = \frac{\partial^2}{\partial\theta\partial x_i}\ \ln g\left(t(x_1, \ldots, x_n);\theta\right) \\ &= \frac{\partial}{\partial x_i}\, t(x_1, \ldots, x_n)\ \frac{\partial^2}{\partial\theta\partial y}\ \ln g(y;\theta)_{|y=t(x_1,\ldots,x_n)}\end{aligned}$$

Pour i et j distincts, on obtient donc :

$$\frac{\dfrac{\partial^2}{\partial\theta\partial x_i}\ \ln f(x_i;\theta)}{\dfrac{\partial^2}{\partial\theta\partial x_j}\ \ln f(x_j;\theta)} = \frac{\dfrac{\partial}{\partial x_i}\, t(x_1, \ldots, x_n)\ \dfrac{\partial^2}{\partial\theta\partial y}\ \ln g(y;\theta)_{|y=t(x_1,\ldots,x_n)}}{\dfrac{\partial}{\partial x_j}\, t(x_1, \ldots, x_n)\ \dfrac{\partial^2}{\partial\theta\partial y}\ \ln g(y;\theta)_{|y=t(x_1,\ldots,x_n)}} = \frac{\dfrac{\partial}{\partial x_i}\, t(x_1, \ldots, x_n)}{\dfrac{\partial}{\partial x_j}\, t(x_1, \ldots, x_n)}$$

qui ne dépend pas de θ. On est donc dans la situation d'une fonction φ telle que $\dfrac{\varphi(x;\theta)}{\varphi(y;\theta)}$ ne dépend pas de θ. Alors forcément $\varphi(x;\theta)$ est de la forme $\varphi(x;\theta) = u(x)v(\theta)$. Par conséquent, on a $\dfrac{\partial^2}{\partial\theta\partial x}\ \ln f(x;\theta) = u(x)v(\theta)$.

D'où $\dfrac{\partial}{\partial\theta}\ \ln f(x;\theta) = a(x)v(\theta) + w(\theta)$ et $\ln f(x;\theta) = a(x)\alpha(\theta) + \beta(\theta) + b(x)$.

Finalement, la densité est bien de la forme $f(x;\theta) = e^{\,a(x)\alpha(\theta) + b(x) + \beta(\theta)}$.

□

Pour finir cette section, appliquons le théorème de Darmois aux lois usuelles.

- *Loi de Bernoulli* $\mathcal{B}(p)$: $a(x) = x$, donc on retrouve le fait que $\sum\limits_{i=1}^{n} x_i$ est une statistique exhaustive.
- *Loi exponentielle* $exp(\lambda)$: $a(x) = x$, donc $\sum\limits_{i=1}^{n} x_i$ est une statistique exhaustive.
- *Loi normale* $\mathcal{N}(m, \sigma^2)$: $a_1(x) = x^2$ et $a_2(x) = x$, donc on retrouve le fait que $\left(\sum\limits_{i=1}^{n} x_i^2, \sum\limits_{i=1}^{n} x_i\right)$ ou $(\bar{x}_n, s_n^2)$ est une statistique exhaustive.
- *Loi de Weibull* $\mathcal{W}(\eta, \beta)$. Elle n'appartient pas à la famille exponentielle, donc il n'y a pas de statistique exhaustive. Cela peut se voir autrement en écrivant la vraisemblance :

$$\mathcal{L}(\eta, \beta; x_1, \ldots, x_n) = \prod_{i=1}^{n} \beta\, \frac{x_i^{\beta-1}}{\eta^\beta}\, e^{-\left(\dfrac{x_i}{\eta}\right)^{\beta}} = \frac{\beta^n}{\eta^{n\beta}} \left[\prod_{i=1}^{n} x_i^{\beta-1}\right] e^{-\dfrac{1}{\eta^\beta}\sum\limits_{i=1}^{n} x_i^\beta}$$

Elle ne peut pas être factorisée sous la forme du théorème de Fisher-Neyman $g\left(t(x_1, \ldots, x_n); \eta, \beta\right)$ $h(x_1, \ldots, x_n)$, sauf si on prend $t(x_1, \ldots, x_n) = (x_1, \ldots, x_n)$. Autrement dit, on ne peut pas résumer l'ensemble des données en conservant la totalité de l'information sur les paramètres.

6.5 Réduction de la variance

Jusqu'à la fin de ce chapitre, on considère le modèle paramétrique $(\mathcal{X}, \mathcal{A}, \{\mathbb{P}_\theta^X\ ;\ \theta \in \Theta\ \subset \mathbb{R}\})$, où le paramètre θ est de dimension 1. Ainsi, la notion de variance minimale a un sens et on va proposer une méthode permettant de déterminer un ESBVM de θ.

Dans un premier temps, le théorème suivant permet, à partir d'un estimateur sans biais, de construire un autre estimateur sans biais de variance inférieure, pour peu qu'il existe une statistique exhaustive.

Théorème 6.5.1. Théorème de Rao-Blackwell. *S'il existe une statistique exhaustive T et un estimateur sans biais $\hat{\theta}$ de θ, alors $Z = \mathbb{E}[\hat{\theta} \,|\, T]$ est un estimateur sans biais de θ, de variance inférieure à celle de $\hat{\theta}$.*

Rappels :

- La loi de Y sachant $[X = x]$ admet pour espérance $\mathbb{E}[Y \mid X = x]$, qui est une fonction de x.
- $\mathbb{E}[Y \mid X]$ est une variable aléatoire fonction de X, dont $\mathbb{E}[Y \mid X = x]$ est une réalisation.
- Formule de l'espérance totale : $\mathbb{E}\left[\mathbb{E}[Y \mid X]\right] = \mathbb{E}[Y]$.

On peut aussi montrer que :

- Pour toute fonction φ mesurable, $\mathbb{E}[\varphi(X) \,|\, X] = \varphi(X)$.
- Pour toute fonction φ mesurable, $\mathbb{E}[\varphi(X)Y \mid X] = \varphi(X)\mathbb{E}[Y \mid X]$.

Démonstration. Comme T est exhaustive, la loi de X sachant T ne dépend pas de θ, donc celle de $\hat{\theta}$ (= $\hat{\theta}(X)$) sachant T non plus. Par conséquent, $\mathbb{E}[\hat{\theta} \,|\, T = t]$ ne dépend pas de θ, donc $z(x) = \mathbb{E}[\hat{\theta} \,|\, T = t(x)]$ est bien une statistique. Ce résultat est indispensable puisque, si Z dépendait de θ, on ne pourrait pas l'utiliser pour estimer θ.

D'après la formule de l'espérance totale, $\mathbb{E}[Z] = \mathbb{E}\left[\mathbb{E}[\hat{\theta} \,|\, T]\right] = \mathbb{E}[\hat{\theta}]$. Donc si $\hat{\theta}$ est un estimateur sans biais de θ, Z est aussi un estimateur sans biais de θ. La variance de $\hat{\theta}$ est :

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\hat{\theta}] &= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2\right] = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] \\
&= \mathbb{E}\left[(\hat{\theta} - Z + Z - \theta)^2\right] \\
&= \mathbb{E}\left[(\hat{\theta} - Z)^2\right] + \mathbb{E}\left[(Z - \theta)^2\right] + 2\mathbb{E}\left[(\hat{\theta} - Z)(Z - \theta)\right].
\end{aligned}$$

Les 3 termes de cette somme vérifient :

1. $\mathbb{E}\left[(\hat{\theta} - Z)^2\right] \geq 0.$
2. $\mathbb{E}\left[(Z - \theta)^2\right] = \mathbb{E}\left[(Z - \mathbb{E}[Z])^2\right] = \mathbb{V}\mathrm{ar}[Z].$
3. $\mathbb{E}\left[(\hat{\theta} - Z)(Z - \theta)\right] = \mathbb{E}\left[(\hat{\theta} - Z)Z\right] - \theta\mathbb{E}[\hat{\theta} - Z] = \mathbb{E}\left[(\hat{\theta} - Z)Z\right]$
   car $\mathbb{E}[\hat{\theta} - Z] = \mathbb{E}[\hat{\theta}] - \mathbb{E}[Z] = \theta - \theta = 0.$

Enfin :

$$\begin{aligned}
\mathbb{E}\left[(\hat{\theta} - Z)Z\right] &= \mathbb{E}\left[\mathbb{E}\left[(\hat{\theta} - Z)Z \,|\, T\right]\right] \quad \text{d'après la formule de l'espérance totale} \\
&= \mathbb{E}\left[\mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta} \,|\, T])\, \mathbb{E}[\hat{\theta} \,|\, T] \,|\, T\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}[\hat{\theta} \,|\, T]\, \mathbb{E}\left[\hat{\theta} - \mathbb{E}[\hat{\theta} \,|\, T] \,|\, T\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}[\hat{\theta} \,|\, T]\, \left[\mathbb{E}[\hat{\theta} \,|\, T] - \mathbb{E}[\hat{\theta} \,|\, T]\right]\right] \\
&= 0.
\end{aligned}$$

D'où $\mathbb{V}\mathrm{ar}[\hat{\theta}] = \mathbb{E}\left[(\hat{\theta} - Z)^2\right] + \mathbb{V}\mathrm{ar}[Z]$, ce qui prouve que $\mathbb{V}\mathrm{ar}[Z] \leq \mathbb{V}\mathrm{ar}[\hat{\theta}]$, d'où le théorème. □

On présente maintenant un exemple qui illustre l'intérêt de théorème de Rao-Blackwell.

Exemple 6.5.1. *Ampoules*. Modèle d'échantillon de loi exponentielle. X_i est la durée de vie d'une ampoule, supposée de loi $\exp(\lambda)$. On souhaite estimer la fiabilité d'une ampoule à l'instant x, c'est-à-dire la probabilité qu'elle fonctionne toujours au bout d'une durée x :

$$R(x) = \mathbb{P}(X_i > x) = 1 - F_{X_i}(x) = e^{-\lambda x}.$$

On sait que l'estimateur de maximum de vraisemblance de λ est $\hat{\lambda}_n = 1/\bar{X}_n = n/\sum\limits_{i=1}^{n} X_i$, donc l'estimateur de maximum de vraisemblance de $R(x)$ est :

$$\hat{R}_n(x) = e^{-\hat{\lambda}_n x} = e^{-nx/\sum\limits_{i=1}^{n} X_i}.$$

On a vu que $\hat{\lambda}$ est un estimateur biaisé de λ et que $\hat{\lambda}'_n = (n-1)/\sum\limits_{i=1}^{n} X_i$ est un estimateur sans biais. On peut donc aussi proposer d'estimer $R(x)$ par $\hat{R}'_n(x) = e^{-(n-1)x/\sum\limits_{i=1}^{n} X_i}$.

Mais le biais de ces estimateurs est difficile à calculer. En effet, étant donné que $\sum\limits_{i=1}^{n} X_i$ est de loi $G(n,\lambda)$, on a par exemple :

$$\mathbb{E}\left[\hat{R}_n(x)\right] = \int_0^{+\infty} e^{-nx/y} \, \frac{\lambda^n}{(n-1)!} \, e^{-\lambda y} \, y^{n-1} \, dy$$

qu'on ne sait pas calculer.

Une autre solution consiste à estimer la probabilité qu'une ampoule fonctionne toujours à l'instant x par le pourcentage d'ampoules observées qui fonctionnent toujours à l'instant x. C'est ce qu'on appelle la fiabilité empirique :

$$\mathbb{R}_n(x) = 1 - \mathbb{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i > x\}}.$$

où $\mathbb{F}_n$ est la fonction de répartition empirique, vue en PS1.

Les propriétés de cet estimateur sont faciles à établir. En effet, les $Y_i = \mathbb{1}_{\{X_i>x\}}$ sont des variables aléatoires indépendantes et de même loi de Bernoulli $\mathcal{B}(\mathbb{P}(Y_i = 1)) = \mathcal{B}(\mathbb{P}(X_i > x)) = \mathcal{B}(R(x))$. La fiabilité empirique n'est autre que la moyenne empirique des Y_i : $\mathbb{R}_n(x) = \bar{Y}_n$. Donc on sait que $\mathbb{R}_n(x)$ est un estimateur sans biais et convergent de $\mathbb{E}[Y_i] = R(x)$:

$$\mathbb{E}\left[\mathbb{R}_n(x)\right] = R(x) \qquad \text{et} \qquad \mathbb{V}\text{ar}\left[\mathbb{R}_n(x)\right] = \frac{\mathbb{V}\text{ar}[Y_i]}{n} = \frac{R(x)\left[1 - R(x)\right]}{n}.$$

On a vu que $t(x) = \sum\limits_{i=1}^{n} x_i$ était une statistique exhaustive pour λ. Par conséquent, le théorème de Rao-Blackwell permet d'affirmer que $Z = \mathbb{E}\left[\mathbb{R}_n(x) \,|\, \sum\limits_{i=1}^{n} X_i\right]$ est un estimateur sans biais de $R(x)$, de variance inférieure à celle de $\mathbb{R}_n(x)$.

Pour calculer Z, on commence par calculer $z(x,t) = \mathbb{E}\left[\mathbb{R}_n(x) | \sum\limits_{i=1}^{n} X_i = t\right]$.

$$\begin{aligned}
z(x,t) &= \mathbb{E}\left[\mathbb{R}_n(x) \,|\, \sum_{i=1}^{n} X_i = t\right] = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{X_j > x\}} \,|\, \sum_{i=1}^{n} X_i = t\right] \\
&= \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\left[\mathbb{1}_{\{X_j > x\}} \,|\, \sum_{i=1}^{n} X_i = t\right] = \mathbb{E}\left[\mathbb{1}_{\{X_1 > x\}} \,|\, \sum_{i=1}^{n} X_i = t\right] \\
&\qquad \text{car les } X_i \text{ sont interchangeables, donc toutes les espérances sont égales} \\
&= \mathbb{P}\big(X_1 > x \,|\, \sum_{i=1}^{n} X_i = t\big) \quad \text{car } \mathbb{E}\left[\mathbb{1}_A\right] = \mathbb{P}(A).
\end{aligned}$$

Comme les X_i sont positives, il est impossible que l'on ait à la fois $X_1 > x$ et $\sum\limits_{i=1}^{n} X_i = t$ quand $t \leq x$. On fera donc le calcul sous l'hypothèse $t > x$ et on rajoutera à la fin l'indicatrice $\mathbb{1}_{\{t>x\}}$. On a :

$$\mathbb{P}\big(X_1 > x \,|\, \sum_{i=1}^{n} X_i = t\big) = \int_x^{+\infty} f_{X_1 \,|\, \sum\limits_{i=1}^{n} X_i = t}(u)\, du$$

avec :

$$f_{X_1 \,|\, \sum\limits_{i=1}^{n} X_i = t}(u) = \frac{f_{(X_1, \sum\limits_{i=1}^{n} X_i)}(u,t)}{f_{\sum\limits_{i=1}^{n} X_i}(t)} = \frac{f_{(X_1, \sum\limits_{i=2}^{n} X_i)}(u,t-u)}{f_{\sum\limits_{i=1}^{n} X_i}(t)}$$

Pour les mêmes raisons que précédemment, le numérateur est nul quand $t \leq u$. Donc dans l'intégrale, la borne sup est en fait t au lieu de $+\infty$.

Pour $u < t$, on a :

$$f_{X_1 \,|\, \sum\limits_{i=1}^{n} X_i = t}(u) = \frac{f_{X_1}(u)\, f_{\sum\limits_{i=2}^{n} X_i}(t-u)}{f_{\sum\limits_{i=1}^{n} X_i}(t)}$$

car X_1 et $\sum\limits_{i=2}^{n} X_i$ sont indépendantes. Comme $\sum\limits_{i=2}^{n} X_i$ est de loi $G(n-1, \lambda)$, on a :

$$f_{X_1 \,|\, \sum\limits_{i=1}^{n} X_i = t}(u) = \frac{\lambda e^{-\lambda u}\, \dfrac{\lambda^{n-1}}{(n-2)!}\, e^{-\lambda(t-u)}\, (t-u)^{n-2}}{\dfrac{\lambda^n}{(n-1)!}\, e^{-\lambda t}\, t^{n-1}} = (n-1)\, \frac{(t-u)^{n-2}}{t^{n-1}}$$

On remarque que cette densité ne dépend pas de λ. C'est parce que $\sum\limits_{i=1}^{n} X_i$ est exhaustive. Ce résultat est indispensable car il faut que Z ne dépende pas de λ. On obtient :

$$\begin{aligned} \mathbb{P}\big(X_1 > x \,|\, \sum_{i=1}^{n} X_i = t\big) &= \int_x^t (n-1)\, \frac{(t-u)^{n-2}}{t^{n-1}}\, du = \frac{1}{t^{n-1}}\, \Big[-(t-u)^{n-1}\Big]_x^t \\ &= \frac{(t-x)^{n-1}}{t^{n-1}} = \Big(1 - \frac{x}{t}\Big)^{n-1}, \text{ avec } x < t. \end{aligned}$$

Donc finalement $z(x,t) = \Big(1 - \dfrac{x}{t}\Big)^{n-1} \mathbb{1}_{\{t>x\}}$ et l'estimateur recherché est :

$$Z = \left(1 - \frac{x}{\sum\limits_{i=1}^{n} X_i}\right)^{n-1} \mathbb{1}_{\{\sum\limits_{i=1}^{n} X_i > x\}}.$$

Autant les estimateurs $\hat{R}_n(x)$, $\hat{R'}_n(x)$ et $\mathbb{R}_n(x)$ semblent naturels, autant celui-ci n'est pas intuitif. Pourtant, c'est le meilleur des quatre, comme on le verra plus tard.

On a vu qu'on pouvait diminuer la variance d'un estimateur sans biais, mais peut-on atteindre la variance minimale ? Pour le déterminer, on doit introduire la notion de statistique complète.

6.6 Complétude

Définition 6.6.1. Une statistique t est **complète** ou **totale** si et seulement si pour toute fonction mesurable φ, on a :

$$\mathbb{E}\left[\varphi(T)\right] = 0,\ \forall \theta \in \Theta \Rightarrow \varphi = 0 \text{ presque partout sur le support de la loi de } T, \text{ c'est-à-dire}$$

partout sauf sur un ensemble de mesure nulle.

Avant de voir à quoi sert cette notion, regardons ce qu'elle donne sur les exemples de référence.

Exemple 6.6.1. *Contrôle de qualité.* $X = (X_1, \dots, X_n)$, où les X_i sont i.i.d. de loi de Bernoulli $\mathcal{B}(p)$. On sait que $t(x_1, \dots, x_n) = \sum_{i=1}^{n} x_i$ est une statistique exhaustive pour p. Est-elle complète ?

On sait que $T = \sum_{i=1}^{n} X_i$ est de loi binomiale $\mathcal{B}(n, p)$, donc :

$$\mathbb{E}[\varphi(T)] = \sum_{k=0}^{n} \varphi(k)\, \mathbb{P}(T = k) = \sum_{k=0}^{n} \varphi(k)\ \binom{n}{k}\ p^k (1-p)^{n-k}.$$

Il faut montrer que

$$\sum_{k=0}^{n} \varphi(k)\ \binom{n}{k}\ p^k (1-p)^{n-k} = 0,\ \forall p \in [0,1]\ \Rightarrow\ \forall k \in \{0, \dots, n\},\ \varphi(k) = 0.$$

En effet, comme le support de T est fini, φ doit être nulle partout sur le support.

Or $\sum_{k=0}^{n} \varphi(k)\ \binom{n}{k}\ p^k (1-p)^{n-k} = (1-p)^n \sum_{k=0}^{n} \varphi(k)\ \binom{n}{k} \left(\frac{p}{1-p}\right)^k$.

Soit $\theta = \frac{p}{1-p}$. Quand p parcourt $[0,1]$, θ parcourt $\mathbb{R}^+$. On a donc :

$$\sum_{k=0}^{n} \varphi(k)\ \binom{n}{k}\ p^k (1-p)^{n-k} = 0,\ \forall p \in [0,1]\ \Rightarrow\ \sum_{k=0}^{n} \varphi(k)\ \binom{n}{k}\ \theta^k = 0,\ \forall \theta \in \mathbb{R}^+.$$

C'est un polynôme de degré n en θ qui est identiquement nul, donc tous ses coefficients sont nuls. Par conséquent, $\forall k \in \{0, \dots, n\}$, $\varphi(k)\ \binom{n}{k} = 0$ et donc $\forall k \in \{0, \dots, n\}$, $\varphi(k) = 0$, ce qui prouve que $t(x_1, \dots, x_n) = \sum_{i=1}^{n} x_i$ est une statistique complète.

Exemple 6.6.2. *Ampoules.* $X = (X_1, \dots, X_n)$, où les X_i sont i.i.d. de loi exponentielle $\exp(\lambda)$. On sait que $t(x_1, \dots, x_n) = \sum_{i=1}^{n} x_i$ est une statistique exhaustive pour λ. Est-elle complète ?

On sait que $T = \sum_{i=1}^{n} X_i$ est de loi gamma $G(n, \lambda)$, donc :

$$\mathbb{E}[\varphi(T)] = \int_0^{+\infty} \varphi(y)\, \frac{\lambda^n}{(n-1)!}\, e^{-\lambda y}\, y^{n-1}\, dy.$$

$$\mathbb{E}[\varphi(T)] = 0,\ \forall \lambda \in \mathbb{R}^+\ \Rightarrow\ \int_0^{+\infty} \varphi(y)\, y^{n-1}\, e^{-\lambda y}\, dy = 0,\ \forall \lambda \in \mathbb{R}^+.$$

La transformée de Laplace d'une fonction f à support positif est la fonction de la variable complexe s définie par $\bar{f}(s) = \int_0^{+\infty} f(y) e^{-sy} dy$. L'intégrale ci-dessus est donc la transformée de Laplace de la fonction $\varphi(y)\, y^{n-1}$ au point λ. Comme la transformée de Laplace est injective, la seule fonction dont la transformée de Laplace est nulle est la fonction nulle.

Donc on a $\forall y \in \mathbb{R}^+$, $\varphi(y)\, y^{n-1} = 0$, d'où $\forall y \in \mathbb{R}^{+*}$, $\varphi(y) = 0$. φ n'est peut-être pas nulle en 0, mais elle est nulle presque partout sur $\mathbb{R}^+$, support de la loi $G(n, \lambda)$. Par conséquent, $t(x_1, \dots, x_n) = \sum_{i=1}^{n} x_i$ est une statistique complète.

6.7 L'estimation sans biais et de variance minimale

Les notions d'exhaustivité et de complétude permettent de trouver un ESBVM de θ à partir d'un estimateur sans biais, grâce au théorème de Lehmann-Scheffé.

Théorème 6.7.1. Théorème de Lehmann-Scheffé. *Si $\hat{\theta}$ est un estimateur sans biais de $\theta \in \mathbb{R}$ et t est une statistique exhaustive et complète, alors $Z = \mathbb{E}[\hat{\theta} \,|\, T]$ est l'unique estimateur sans biais et de variance minimale de θ.*

Démonstration. D'après le théorème de Rao-Blackwell, si un estimateur sans biais $\hat{\theta}$ n'est pas fonction de la statistique exhaustive T, on peut toujours trouver un autre estimateur sans biais de θ, de variance inférieure, qui soit fonction de $T : Z = \mathbb{E}[\hat{\theta} \,|\, T]$. Donc un ESBVM est forcément fonction de T.

Supposons qu'il existe 2 estimateurs sans biais fonction de T, $\hat{\theta}_1(T)$ et $\hat{\theta}_2(T)$.

$$\mathbb{E}\left[\hat{\theta}_1(T)\right] = \mathbb{E}\left[\hat{\theta}_2(T)\right] = \theta \text{ donc } \forall \theta \in \Theta, \, \mathbb{E}\left[\hat{\theta}_1(T) - \hat{\theta}_2(T)\right] = \mathbb{E}\left[(\hat{\theta}_1 - \hat{\theta}_2)(T)\right] = 0.$$

Comme t est complète, on en déduit que $\hat{\theta}_1 - \hat{\theta}_2 = 0$ presque partout, d'où $\hat{\theta}_1 = \hat{\theta}_2$ presque partout. Il n'existe donc qu'un seul estimateur sans biais fonction de T et cet estimateur est de variance minimale. La complétude ne sert qu'à garantir l'unicité de l'ESBVM. □

Corollaire 6.7.1. *Pour trouver un estimateur optimal, il suffit de trouver un estimateur sans biais fonction mesurable d'une statistique exhaustive et complète.*

Exemple 6.7.1. *Contrôle de qualité.* $\hat{p}_n = \bar{X}_n = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i$ est un estimateur sans biais de p, fonction de la statistique exhaustive et complète $\sum\limits_{i=1}^{n} X_i$, donc c'est l'ESBVM de p. On retrouve le résultat déjà montré d'une manière différente dans l'exemple 6.1.1.

Exemple 6.7.2. *Ampoules.* L'estimateur de maximum de vraisemblance de λ est $\hat{\lambda}_n = n / \sum\limits_{i=1}^{n} X_i$. On a vu qu'il était biaisé et que $\hat{\lambda}'_n = (n-1) / \sum\limits_{i=1}^{n} X_i$ était sans biais. $\hat{\lambda}'_n$ est sans biais et fonction de la statistique exhaustive et complète $\sum\limits_{i=1}^{n} X_i$, donc c'est l'ESBVM de λ. D'après un résultat vu en TD, cet estimateur n'est pas efficace (sa variance est strictement supérieure à la borne de Cramer-Rao). Donc le théorème de Lehmann-Scheffé est le seul moyen de prouver l'optimalité de cet estimateur.

Proposition 6.7.1. *S'il existe un estimateur sans biais de $\varphi(\theta)$ fonction mesurable d'une statistique exhaustive et complète, alors c'est l'ESBVM de $\varphi(\theta)$.*

Exemple 6.7.3. *Ampoules.* On a vu que $Z = \left(1 - \dfrac{x}{\sum\limits_{i=1}^{n} X_i}\right)^{n-1} \mathbb{1}_{\{\sum\limits_{i=1}^{n} X_i > x\}}$ est un estimateur sans biais de $R(x) = e^{-\lambda x}$. Comme il est fonction de la statistique exhaustive et complète $\sum\limits_{i=1}^{n} X_i$, cela signifie que Z est l'ESBVM de $R(x)$. $\mathbb{R}_n(x)$ est aussi un estimateur sans biais de $R(x)$, mais comme il n'est pas fonction de $\sum\limits_{i=1}^{n} X_i$, ce n'est pas l'ESBVM.

Théorème 6.7.2. *Dans un modèle d'échantillon où la loi des observations appartient à la famille exponentielle, si $\alpha(\theta)$ est bijective, alors la statistique exhaustive $\sum\limits_{i=1}^{n} a(x_i)$ est complète.*

Ce théorème permet de retrouver directement que $\sum\limits_{i=1}^{n} x_i$ est complète dans les exemples du contrôle de qualité et des ampoules.

Les résultats précédents ont été établis pour un paramètre θ de dimension 1. Pour une dimension $d > 1$, la notion d'estimateur de variance minimale n'a plus de sens. En effet, un estimateur de T de dimension $d > 1$ n'a pas de variance mais a une matrice de covariance.

En dimension 1, on a vu que l'erreur quadratique moyenne (EQM) d'un estimateur est la somme de sa variance et du carré de son biais. Pour un estimateur sans biais, minimiser la variance est donc équivalent à minimiser l'EQM. En dimension d, l'EQM s'écrit :

$$EQM[T] = \mathbb{E}\left[||T-\theta||^2\right] = \sum_{j=1}^{d} \mathbb{E}\left[\left(T_j - \theta_j\right)^2\right]$$

Si T est un estimateur sans biais de θ, $\mathbb{E}[T] = \theta$, donc $\forall j, \mathbb{E}[T_j] = \theta_j$. Alors $EQM[T] = \sum\limits_{j=1}^{d} \mathbb{V}\text{ar}[T_j]$.

Par analogie avec la dimension 1, en dimension $d > 1$, on cherchera des estimateurs sans biais d'EQM minimale. Les théorèmes de Rao-Blackwell et Lehmann-Scheffé se généralisent alors de la façon suivante : on réduit l'EQM d'un estimateur sans biais en prenant son espérance conditionnellement à une statistique exhaustive et on a l'EQM minimale si cette statistique est complète.

Il existe également un équivalent du théorème 6.7.2 pour $d > 1$. Cela permet par exemple de montrer que, pour un échantillon de loi normale, la statistique exhaustive $\left(\sum\limits_{i=1}^{n} x_i^2, \sum\limits_{i=1}^{n} x_i\right)$ est complète. Alors, $(\bar{X}_n, S'^2_n)^T$ est un estimateur sans biais de $(m, \sigma^2)^T$ fonction de la statistique exhaustive et complète, donc $\bar{X}_n$ et S'^2_n sont les ESBVM de m et de σ^2.

Chapitre 7

Maximum de vraisemblance

On se place dans un modèle paramétrique $(\mathcal{X}, \mathcal{A}, \{\mathbb{P}_\theta^X ; \theta \in \Theta \subset \mathbb{R}^d\})$ et on souhaite estimer θ. Dans ce chapitre nous allons étudier les propriétés des estimateurs de maximum de vraisemblance et montrer qu'ils sont asymptotiquement optimaux. Les résultats établis permettront en particulier de construire des intervalles de confiance asymptotiques pour les paramètres du modèle sous-jacent.

7.1 Propriétés des estimateurs de maximum de vraisemblance

Rappelons que si la fonction de vraisemblance $\mathcal{L}(\theta; x)$ admet un maximum unique au point $\hat{\theta}(x)$, alors l'application $x \mapsto \hat{\theta}(x)$ est appelée statistique de maximum de vraisemblance et $\hat{\theta}(X)$ est l'estimateur de maximum de vraisemblance (EMV) de θ. Dans la suite, on notera plus simplement $\hat{\theta}$ cet estimateur. On a donc :

$$\hat{\theta} = \arg\max_\theta \mathcal{L}(\theta; X).$$

Comme d'habitude, on préfèrera maximiser le logarithme de la vraisemblance :

$$\hat{\theta} = \arg\max_\theta \ln \mathcal{L}(\theta; X).$$

Dans la plupart des cas, on maximisera la log-vraisemblance en annulant sa dérivée par rapport à chaque composante de θ. Mais on a vu (voir le cas de la loi uniforme) que cette méthode ne fonctionnait pas toujours. Nous allons nous placer dans ce chapitre dans le cas où cette méthode va fonctionner. Il faut pour cela faire les hypothèses 6.2.1 (dérivabilité, intégration,...) indispensables pour définir la matrice d'information. Dans ces conditions, l'EMV $\hat{\theta}$ est solution du système des équations de vraisemblance :

$$\forall j \in \{1, \ldots, d\}, \quad \frac{\partial}{\partial \theta_j} \ln \mathcal{L}(\theta; X) = 0.$$

Mais comme le score est défini par $Z(\theta; X) = \nabla_\theta \ln \mathcal{L}(\theta; X)$, $\hat{\theta}$ est finalement la valeur de θ qui annule le score :

$$Z(\hat{\theta}; X) = 0.$$

Pour assurer la convergence presque sûre du théorème 7.1.1, on a besoin de quelques conditions supplémentaires. Ces conditions étant vérifiées la plupart du temps, on supposera que ce sera le cas dans la suite.

Un estimateur de maximum de vraisemblance n'est pas forcément sans biais, ni de variance minimale, ni efficace. Mais il possède d'excellentes propriétés asymptotiques, que nous allons établir pour un paramètre de dimension d quelconque. Nous nous intéressons ici uniquement aux modèles d'échantillon, mais il existe des résultats analogues pour de nombreux autres modèles. Pour un échantillon de taille n, l'EMV sera noté $\hat{\theta}_n$, le score $Z^{(n)}(\theta; X)$ et la matrice d'information $\mathcal{I}_n(\theta) = n\mathcal{I}_1(\theta)$.

Théorème 7.1.1. *Dans un modèle paramétrique d'échantillon* $(\mathcal{X}, \mathcal{A}, \{\mathbb{P}_\theta^X ; \theta \in \Theta \subset \mathbb{R}^d\})^n$ *vérifiant les hypothèses annoncées, on a :*

- *$\hat{\theta}_n$ converge presque sûrement vers θ.*
- $\sqrt{n}\,(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}_d\left(0, \mathcal{I}_1^{-1}(\theta)\right)$, *où $\mathcal{N}_d$ est la loi normale dans $\mathbb{R}^d$.*

Interprétation de la convergence en loi :

- $\mathbb{E}\left[\sqrt{n}\,(\hat{\theta}_n - \theta)\right]$ tend vers 0, donc $\mathbb{E}[\hat{\theta}_n]$ tend vers θ. Par conséquent, l'EMV est asymptotiquement sans biais.
- $K_{\sqrt{n}\,(\hat{\theta}_n-\theta)} = K_{\sqrt{n}\,\hat{\theta}_n} = nK_{\hat{\theta}_n}$ tend vers $\mathcal{I}_1^{-1}(\theta)$. Donc la matrice de covariance de $\hat{\theta}_n$, $K_{\hat{\theta}_n}$, est asymptotiquement équivalente à $\mathcal{I}_1^{-1}(\theta)/n = [n\mathcal{I}_1]^{-1}(\theta) = \mathcal{I}_n^{-1}(\theta)$, qui est la borne de Cramer-Rao. Par conséquent, l'EMV est asymptotiquement efficace.
- Comme l'EMV est asymptotiquement sans biais et efficace, il est asymptotiquement optimal.
- L'EMV est asymptotiquement gaussien.
- On dit aussi que la **loi asymptotique** de l'EMV $\hat{\theta}_n$ est la loi $\mathcal{N}_d\left(\theta, \mathcal{I}_n^{-1}(\theta)\right)$.
- La vitesse de convergence de $\hat{\theta}_n$ vers θ est $1/\sqrt{n}$, ce qui signifie que la variance de chaque composante de $\hat{\theta}_n$ tend vers 0 comme $1/n$. La proposition 6.2.2 montre que tout estimateur sans biais T de θ est tel que $\forall i \in \{1,\dots,d\}$, $\mathbb{V}\mathrm{ar}[T_i] \geq \mathcal{I}_n^{-1}(\theta)_{ii} = \mathcal{I}_1^{-1}(\theta)_{ii}/n$, donc aucun estimateur sans biais ne peut converger plus vite que $1/\sqrt{n}$.

Démonstration. Nous allons montrer le résultat de convergence en loi pour un paramètre réel ($d = 1$). Alors la quantité d'information est simplement un réel $\mathcal{I}_n(\theta) = n\mathcal{I}_1(\theta)$.

Par commodité d'écriture, on suppose que la loi sous-jacente est continue, de densité f. Alors la vraisemblance s'écrit $\mathcal{L}(\theta; x) = \mathcal{L}(\theta; x_1, \dots, x_n) = \prod_{i=1}^{n} f(x_i;\theta)$ et le score est :

$$Z^{(n)}(\theta; X) = \frac{\partial}{\partial\theta}\ln \mathcal{L}(\theta; X) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta}\ln f(X_i;\theta).$$

On a déjà vu que $\mathbb{E}\left[Z^{(n)}(\theta; X)\right] = 0$ et :

$$\mathcal{I}_n(\theta) = \mathbb{V}\mathrm{ar}\left[Z^{(n)}(\theta; X)\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ln \mathcal{L}(\theta; X)\right] = -\mathbb{E}\left[\frac{\partial}{\partial\theta} Z^{(n)}(\theta; X)\right].$$

En particulier, $\mathcal{I}_1(\theta) = \mathbb{V}\mathrm{ar}\left[\frac{\partial}{\partial\theta}\ln f(X_1;\theta)\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f(X_1;\theta)\right]$.

Les variables aléatoires $\frac{\partial}{\partial\theta}\ln f(X_i;\theta)$ sont indépendantes, de même loi, centrées et de variance $\mathcal{I}_1(\theta)$. Donc le théorème central limite entraine que

$$\sqrt{n}\,\frac{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f(X_i;\theta) - 0}{\sqrt{\mathcal{I}_1(\theta)}} = \frac{Z^{(n)}(\theta; X)}{\sqrt{n\mathcal{I}_1(\theta)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Et par conséquent :

$$\frac{1}{\sqrt{\mathcal{I}_1(\theta)}}\,\frac{Z^{(n)}(\theta; X)}{\sqrt{n\mathcal{I}_1(\theta)}} = \frac{Z^{(n)}(\theta; X)}{\sqrt{n}\mathcal{I}_1(\theta)} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{\mathcal{I}_1(\theta)}\right). \tag{7.1.1}$$

Dans la suite, pour éviter des confusions d'écriture, on va noter θ_0 la vraie valeur (inconnue) du paramètre θ. Le théorème des accroissements finis permet d'écrire qu'il existe un θ'_n entre $\hat{\theta}_n$ et θ_0 tel que :

$$Z^{(n)}(\hat{\theta}_n; X) = Z^{(n)}(\theta_0; X) + (\hat{\theta}_n - \theta_0)\,\frac{\partial}{\partial\theta} Z^{(n)}(\theta; X)_{|\theta'_n}.$$

Or $Z^{(n)}(\hat{\theta}_n; X) = 0$. Multiplions par $1/\sqrt{n}$.

$$\frac{1}{\sqrt{n}} Z^{(n)}(\theta_0; X) + \frac{1}{\sqrt{n}}(\hat{\theta}_n - \theta_0)\,\frac{\partial}{\partial\theta} Z^{(n)}(\theta; X)_{|\theta'_n} = 0$$

$$\text{ou } \frac{1}{\sqrt{n}}\, Z^{(n)}(\theta_0;X) + \sqrt{n}\,(\hat{\theta}_n - \theta_0)\, \frac{\partial}{\partial\theta}\, \frac{1}{n}\, Z^{(n)}(\theta;X)_{\big|\theta'_n} = 0.$$

Donc $\sqrt{n}\,(\hat{\theta}_n - \theta_0) = -\dfrac{\dfrac{1}{\sqrt{n}}\, Z^{(n)}(\theta_0;X)}{\dfrac{\partial}{\partial\theta}\, \dfrac{1}{n}\, Z^{(n)}(\theta;X)_{\big|\theta'_n}}$. On décompose le dénominateur :

$$\begin{aligned}
\frac{\partial}{\partial\theta}\, \frac{1}{n}\, Z^{(n)}(\theta;X)_{\big|\theta'_n} &= \frac{\partial}{\partial\theta}\, \frac{1}{n}\, Z^{(n)}(\theta;X)_{\big|\theta'_n} - \frac{\partial}{\partial\theta}\, \frac{1}{n}\, Z^{(n)}(\theta;X)_{\big|\theta_0} + \frac{\partial}{\partial\theta}\, \frac{1}{n}\, Z^{(n)}(\theta;X)_{\big|\theta_0} + \mathcal{I}_1(\theta_0) - \mathcal{I}_1(\theta_0) \\
&= A_n + B_n - \mathcal{I}_1(\theta_0)
\end{aligned}$$

On a :

$$A_n = \frac{\partial}{\partial\theta}\, \frac{1}{n}\, Z^{(n)}(\theta;X)_{\big|\theta'_n} - \frac{\partial}{\partial\theta}\, \frac{1}{n}\, Z^{(n)}(\theta;X)_{\big|\theta_0}.$$

Puisque $\hat{\theta}_n \xrightarrow{p.s.} \theta_0$ et $\theta'_n \in \Big]\min(\hat{\theta}_n,\theta_0), \max(\hat{\theta}_n,\theta_0)\Big[$, on a forcément $\theta'_n \xrightarrow{p.s.} \theta_0$, donc $A_n \xrightarrow{p.s.} 0$.
Par ailleurs :

$$\begin{aligned}
B_n &= \frac{\partial}{\partial\theta}\, \frac{1}{n}\, Z^{(n)}(\theta;X)_{\big|\theta_0} + \mathcal{I}_1(\theta_0) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2}\, \ln f(X_i;\theta)_{\big|\theta_0} - \mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\, \ln f(X_1;\theta)\right]_{\big|\theta_0}.
\end{aligned}$$

Comme les X_i sont indépendantes et de même loi, la loi forte des grands nombres permet d'affirmer que :

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2}\, \ln f(X_i;\theta)_{\big|\theta_0} \xrightarrow{p.s.} \mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\, \ln f(X_1;\theta)\right]_{\big|\theta_0}$$

donc $B_n \xrightarrow{p.s.} 0$.

Par conséquent, $\dfrac{\partial}{\partial\theta}\, \dfrac{1}{n}\, Z^{(n)}(\theta;X)_{\big|\theta'_n} = A_n + B_n - \mathcal{I}_1(\theta_0) \xrightarrow{p.s.} -\mathcal{I}_1(\theta_0)$.

Pour terminer, on va utiliser le théorème 4.2.6 de Slutsky avec $U_n = \dfrac{1}{\sqrt{n}} Z^{(n)}(\theta_0;X)$ (qui converge en loi vers la loi $\mathcal{N}(0, \mathcal{I}_1(\theta_0))$, $V_n = \mathcal{I}_1(\theta_0) - A_n - B_n$ (qui converge presque sûrement, donc aussi en probabilité, vers $c = \mathcal{I}_1(\theta_0)$), et $g(U_n,V_n) = U_n/V_n$.

Alors $\sqrt{n}\,(\hat{\theta}_n - \theta_0) = \dfrac{Z^{(n)}(\theta_0;X)}{\sqrt{n}\,[\mathcal{I}_1(\theta_0) - A_n - B_n]} = g(U_n,V_n)$ a même limite en loi que $g(U_n,c) = \dfrac{Z^{(n)}(\theta_0;X)}{\sqrt{n}\mathcal{I}_1(\theta_0)}$, à savoir d'après l'équation (7.1.1), la loi $\mathcal{N}\left(0, \dfrac{1}{\mathcal{I}_1(\theta_0)}\right)$. D'où le résultat. □

Si au lieu d'estimer directement θ, on veut estimer une fonction de θ, on sait que $\varphi(\hat{\theta}_n)$ est l'estimateur de maximum de vraisemblance de $\varphi(\theta)$. Les propriétés de cet estimateur sont données par le théorème suivant. Il porte le nom de méthode delta car ce résultat fournit une méthode pour construire des intervalles de confiance asymptotiques.

Théorème 7.1.2. Méthode delta. *Si φ est une fonction de $\mathbb{R}^d$ dans $\mathbb{R}^q$ dérivable par rapport à chaque composante de θ, on a :*

$$\sqrt{n}\left[\varphi(\hat{\theta}_n) - \varphi(\theta)\right] \xrightarrow{\mathcal{L}} \mathcal{N}_q\left(0, \Delta(\theta)\mathcal{I}_1^{-1}(\theta)\Delta(\theta)^T\right)$$

où $\Delta(\theta)$ est la matrice de terme général $\Delta_{ij}(\theta) = \dfrac{\partial}{\partial\theta_j}\, \varphi_i(\theta)$, $1 \leq i \leq q$, $1 \leq j \leq d$.

Démonstration. On fait la démonstration pour $d = q = 1$. Dans ce cas, $\Delta(\theta) = \varphi'(\theta)$, donc le résultat s'écrit :

$$\sqrt{n}\left[\varphi(\hat{\theta}_n) - \varphi(\theta)\right] \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\varphi'(\theta)^2}{\mathcal{I}_1(\theta)}\right)$$

On le montre facilement à l'aide du théorème des accroissements finis. Il existe θ'_n dans $\left]\min(\hat{\theta}_n, \theta), \max(\hat{\theta}_n, \theta)\right[$ tel que :

$$\varphi(\hat{\theta}_n) = \varphi(\theta) + (\hat{\theta}_n - \theta)\varphi'(\theta'_n).$$

Donc $\sqrt{n}\left[\varphi(\hat{\theta}_n) - \varphi(\theta)\right] = \sqrt{n}(\hat{\theta}_n - \theta)\varphi'(\theta'_n)$. Comme $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{\mathcal{I}_1(\theta)}\right)$ et $\varphi'(\theta'_n) \xrightarrow{p.s.} \varphi'(\theta)$, on a bien le résultat ci-dessus.

L'EMV de $\varphi(\theta)$ est donc asymptotiquement sans biais et sa variance asymptotique est

$$\mathbb{V}\mathrm{ar}_{as}[\varphi(\hat{\theta}_n)] = \frac{\varphi'(\theta)^2}{n\mathcal{I}_1(\theta)}$$

□

Le fait que l'EMV soit asymptotiquement sans biais et efficace fait que, si on a beaucoup de données, on est pratiquement certains que la méthode du maximum de vraisemblance est la meilleure méthode d'estimation possible. C'est pourquoi cette méthode est considérée comme globalement la meilleure et est utilisée de préférence à toute autre méthode, y compris celle des moments.

Exemple des ampoules. $X_1, \ldots, X_n$ sont indépendantes et de même loi $\exp(\lambda)$. L'information de Fisher a été calculée en TD et vaut $\mathcal{I}_1(\lambda) = 1/\lambda^2$.

L'EMV de λ est $\hat{\lambda}_n = \dfrac{1}{\overline{X}_n} = \dfrac{n}{\sum\limits_{i=1}^{n} X_i}$. Le résultat asymptotique sur l'EMV s'écrit donc :

$$\sqrt{n}(\hat{\lambda}_n - \lambda) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \mathcal{I}_1^{-1}(\lambda)\right) = \mathcal{N}(0, \lambda^2).$$

Donc la variance asymptotique de $\hat{\lambda}_n$ est $\mathbb{V}\mathrm{ar}_{as}(\hat{\lambda}_n) = \lambda^2/n$. Or en PS1, on a vu que $\mathbb{V}\mathrm{ar}(\hat{\lambda}_n) = \dfrac{n^2\lambda^2}{(n-1)^2(n-2)}$, qui est bien équivalent à λ^2/n.

L'EMV de $R(x) = \varphi(\lambda) = e^{-\lambda x}$ est $\hat{R}_n(x) = e^{-\hat{\lambda}_n x}$. On a vu dans le chapitre 6 qu'on ne pouvait pas calculer son biais et sa variance pour n fini. Mais la méthode delta montre que $\hat{R}_n(x)$ est asymptotiquement sans biais et que sa variance asymptotique est :

$$\mathbb{V}\mathrm{ar}_{as}\left(\hat{R}_n(x)\right) = \frac{\varphi'(\lambda)^2}{n\mathcal{I}_1(\lambda)} = \frac{x^2 e^{-2\lambda x}}{n/\lambda^2} = \frac{\lambda^2 x^2}{n}\, e^{-2\lambda x}.$$

7.2 Intervalles de confiance asymptotiques

On a vu en PS1 qu'un intervalle de confiance de seuil α pour un paramètre $\theta \in \mathbb{R}$, est un intervalle aléatoire $[Y, Z]$ qui a une probabilité $1 - \alpha$ de contenir θ :

Définition 7.2.1. $[Y, Z]$ est un **intervalle de confiance** (exact) de seuil α pour θ si et seulement si :

$$\mathbb{P}\left(\theta \in [Y, Z]\right) = 1 - \alpha.$$

La meilleure façon de déterminer un intervalle de confiance est de trouver une **fonction pivotale** (ou pivot), c'est-à-dire une fonction des observations et du paramètre dont la loi de probabilité ne dépend pas du paramètre. Mais il n'est pas forcément facile de trouver une telle fonction. Nous allons voir dans cette section que les propriétés asymptotiques de l'estimateur de maximum de vraisemblance permettent de déterminer assez facilement des intervalles de confiance asymptotiques pour des fonctions presque quelconques des paramètres.

On se place ici dans un modèle paramétrique d'échantillon de taille n. Pour mettre en évidence la taille de l'échantillon, on notera $[Y_n, Z_n]$ un intervalle de confiance pour θ.

Définition 7.2.2. $[Y_n, Z_n]$ est un **intervalle de confiance asymptotique** de seuil α pour θ si et seulement si :

$$\lim_{n \to +\infty} \mathbb{P}\left(\theta \in [Y_n, Z_n]\right) = 1 - \alpha.$$

Dans la pratique, si on sait calculer un intervalle de confiance exact, on n'a pas besoin de calculer un intervalle de confiance asymptotique. Mais quand on ne sait pas calculer un intervalle de confiance exact, on utilise un intervalle de confiance asymptotique : si n est suffisamment grand, $\mathbb{P}\left(\theta \in [Y_n, Z_n]\right)$ ne devrait pas être trop éloigné de $1 - \alpha$.

Nous allons d'abord déterminer des intervalles de confiance asymptotiques pour un paramètre $\theta \in \mathbb{R}$. Pour le cas où $\theta = (\theta_1, \ldots, \theta_d)^T \in \mathbb{R}^d$ avec $d > 1$, la notion d'intervalle de confiance se généralise en celle de **région de confiance** : région E de $\mathbb{R}^d$ telle que $\mathbb{P}(\theta \in E) = 1 - \alpha$. Mais on se contente la plupart du temps de déterminer des intervalles de confiance, exacts ou asymptotiques, pour chaque composante θ_j de θ. C'est ce que nous ferons à la fin de cette section.

7.2.1 Cas d'un paramètre réel

Si $\theta \in \mathbb{R}$, $\mathcal{I}_1(\theta)$ est un réel et le résultat asymptotique sur l'EMV s'écrit : $\sqrt{n}\,(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \dfrac{1}{\mathcal{I}_1(\theta)}\right)$ ou $\sqrt{n\mathcal{I}_1(\theta)}\,(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$. Le terme $\sqrt{n\mathcal{I}_1(\theta)}\,(\hat{\theta}_n - \theta)$ est une fonction pivotale asymptotique : fonction de θ et des observations (par l'intermédiaire de $\hat{\theta}_n = \hat{\theta}_n(X)$), dont la loi asymptotique ne dépend pas de θ.

On note $u_\alpha = F^{-1}_{\mathcal{N}(0,1)}\left(1 - \frac{\alpha}{2}\right)$ le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0,1)$, qu'on peut obtenir en `R` avec la commande `qnorm(1-alpha/2)` ou à l'aide des tables du chapitre 9. Si Y est de loi $\mathcal{N}(0,1)$, $\mathbb{P}(-u_\alpha \leq Y \leq u_\alpha) = 1 - \alpha$. On a donc :

$$\lim_{n \to +\infty} P\left(-u_\alpha \leq \sqrt{n\mathcal{I}_1(\theta)}\,(\hat{\theta}_n - \theta) \leq +u_\alpha\right) = \lim_{n \to +\infty} P\left(\hat{\theta}_n - \frac{u_\alpha}{\sqrt{n\mathcal{I}_1(\theta)}} \leq \theta \leq \hat{\theta}_n + \frac{u_\alpha}{\sqrt{n\mathcal{I}_1(\theta)}}\right) = 1 - \alpha$$

Donc $\left[\hat{\theta}_n - \dfrac{u_\alpha}{\sqrt{n\mathcal{I}_1(\theta)}}, \hat{\theta}_n + \dfrac{u_\alpha}{\sqrt{n\mathcal{I}_1(\theta)}}\right]$ est un intervalle de confiance asymptotique de seuil α pour θ. Mais cet intervalle est inutilisable à cause du terme $\mathcal{I}_1(\theta)$ qui est inconnu. L'idée naturelle est de le remplacer par $\mathcal{I}_1(\hat{\theta}_n)$. Pour savoir quel est l'impact de cette transformation, il faut utiliser le théorème 4.2.6 de Slutsky.

On pose $U_n = \sqrt{n}\,(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \dfrac{1}{\mathcal{I}_1(\theta)}\right)$.

On a fait l'hypothèse que le modèle statistique sous-jacent vérifiait les conditions permettant d'assurer que $\hat{\theta}_n \xrightarrow{p.s.} \theta$. On a donc également $\sqrt{\mathcal{I}_1(\hat{\theta}_n)} \xrightarrow{p.s.} \sqrt{\mathcal{I}_1(\theta)}$. Comme la convergence presque sûre entraîne la convergence en probabilité, on a aussi $\sqrt{\mathcal{I}_1(\hat{\theta}_n)} \xrightarrow{P} \sqrt{\mathcal{I}_1(\theta)}$.

Soit $g(u,v) = uv$, $V_n = \sqrt{\mathcal{I}_1(\hat{\theta}_n)}$ et $c = \sqrt{\mathcal{I}_1(\theta)}$. Le théorème de Slutsky permet d'écrire que $g(U_n, V_n) = \sqrt{n\mathcal{I}_1(\hat{\theta}_n)}\,(\hat{\theta}_n - \theta)$ a même limite en loi que $g(U_n, c) = \sqrt{n\mathcal{I}_1(\theta)}\,(\hat{\theta}_n - \theta)$, donc $\sqrt{n\mathcal{I}_1(\hat{\theta}_n)}\,(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$. Alors, en appliquant la même démarche que précédemment, on obtient la propriété suivante.

Proposition 7.2.1. *Un intervalle de confiance asymptotique de seuil α pour θ est :*

$$\left[\hat{\theta}_n - \frac{u_\alpha}{\sqrt{n\mathcal{I}_1(\hat{\theta}_n)}}, \hat{\theta}_n + \frac{u_\alpha}{\sqrt{n\mathcal{I}_1(\hat{\theta}_n)}}\right].$$

Cet intervalle a donc la même expression que l'intervalle vu plus haut, en remplaçant θ (inconnu) par $\hat{\theta}_n$ (observé).

Exemple 1 : contrôle de qualité. $X_1, \ldots, X_n$ sont indépendantes et de même loi $\mathcal{B}(p)$. On a vu dans le chapitre 7 que $\mathcal{I}_n(p) = n\mathcal{I}_1(p) = \dfrac{n}{p(1-p)}$. Donc un intervalle de confiance asymptotique de seuil α pour p est :

$$\left[\hat{p}_n - u_\alpha\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \hat{p}_n + u_\alpha\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right].$$

Exemple 2 : ampoules. $X_1, \ldots, X_n$ sont indépendantes et de même loi $\exp(\lambda)$. $\mathcal{I}_n(\lambda) = n\mathcal{I}_1(\lambda) = \dfrac{n}{\lambda^2}$. Donc un intervalle de confiance asymptotique de seuil α pour λ est :

$$\left[\hat{\lambda}_n - u_\alpha\frac{\hat{\lambda}_n}{\sqrt{n}}, \hat{\lambda}_n + u_\alpha\frac{\hat{\lambda}_n}{\sqrt{n}}\right] = \left[\hat{\lambda}_n\left(1 - \frac{u_\alpha}{\sqrt{n}}\right), \hat{\lambda}_n\left(1 + \frac{u_\alpha}{\sqrt{n}}\right)\right].$$

On a vue en PS1 qu'un intervalle de confiance exact est :

$$\left[\hat{\lambda}_n\,\frac{F^{-1}_{G(n,1)}(\alpha/2)}{n}, \hat{\lambda}_n\,\frac{F^{-1}_{G(n,1)}(1-\alpha/2)}{n}\right].$$

Pour n grand, les deux intervalles de confiance sont équivalents.

Intéressons-nous maintenant à des intervalles de confiance asymptotiques pour une fonction $\varphi(\theta)$ du paramètre θ, où $\theta \in \mathbb{R}$ et φ est continue et dérivable. Le résultat de la méthode delta s'écrit :

$$\sqrt{n}\left[\varphi(\hat{\theta}_n) - \varphi(\theta)\right] \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\varphi'(\theta)^2}{\mathcal{I}_1(\theta)}\right)$$

ou :

$$\frac{\sqrt{n\mathcal{I}_1(\theta)}}{|\varphi'(\theta)|}\left[\varphi(\hat{\theta}_n) - \varphi(\theta)\right] \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

On peut encore appliquer le théorème de Slutsky et on obtient le résultat suivant.

Proposition 7.2.2. *Un intervalle de confiance asymptotique de seuil α pour $\varphi(\theta)$ est :*

$$\left[\varphi(\hat{\theta}_n) - u_\alpha\,\frac{|\varphi'(\hat{\theta}_n)|}{\sqrt{n\mathcal{I}_1(\hat{\theta}_n)}}, \varphi(\hat{\theta}_n) + u_\alpha\,\frac{|\varphi'(\hat{\theta}_n)|}{\sqrt{n\mathcal{I}_1(\hat{\theta}_n)}}\right].$$

Exemple des ampoules. L'estimateur de maximum de vraisemblance de $R(x) = \varphi(\lambda) = e^{-\lambda x}$ est $e^{-\hat{\lambda}_n x}$. On a vu que

$$\frac{\varphi'(\lambda)^2}{n\mathcal{I}_1(\lambda)} = \frac{\lambda^2 x^2}{n}\,e^{-2\lambda x}.$$

Donc un intervalle de confiance asymptotique de seuil α pour $R(x)$ est :

$$\left[e^{-\hat{\lambda}_n x} - u_\alpha\,\frac{\hat{\lambda}_n x}{\sqrt{n}}\,e^{-\hat{\lambda}_n x}, e^{-\hat{\lambda}_n x} + u_\alpha\,\frac{\hat{\lambda}_n x}{\sqrt{n}}\,e^{-\hat{\lambda}_n x}\right].$$

7.2.2 Cas d'un paramètre vectoriel

Pour $\theta \in \mathbb{R}^d$, le résultat asymptotique sur l'EMV s'écrit :

$$\sqrt{n}\,(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}_d\left(0, \mathcal{I}_1^{-1}(\theta)\right).$$

$\mathcal{I}_1(\theta)$ est une matrice symétrique définie positive, donc on peut prendre son unique racine carrée définie positive $\mathcal{I}_1^{1/2}(\theta)$ et écrire :

$$\sqrt{n}\,\mathcal{I}_1^{1/2}(\theta)\,(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}_d\,(0, Id)\,.$$

où Id est la matrice identité. Sous des conditions de régularité (continuité des composantes de $\mathcal{I}_1(\theta)$ par rapport à chaque composante de θ), on peut appliquer une version vectorielle du théorème de Slutsky et on obtient :

$$\sqrt{n}\,\mathcal{I}_1^{1/2}(\hat{\theta}_n)\,(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}_d\,(0, Id)$$

ce qui permet de donner des intervalles de confiance asymptotiques pour chaque composante de θ.

De même, le résultat de la méthode delta s'écrit :

$$\sqrt{n}\left[\varphi(\hat{\theta}_n) - \varphi(\theta)\right] \xrightarrow{\mathcal{L}} \mathcal{N}_q\left(0, \Delta(\theta)\mathcal{I}_1^{-1}(\theta)\Delta(\theta)^T\right)$$

ou :

$$\sqrt{n}\,\left[\Delta(\theta)\mathcal{I}_1^{-1}(\theta)\Delta(\theta)^T\right]^{-1/2}\,\left[\varphi(\hat{\theta}_n) - \varphi(\theta)\right] \xrightarrow{\mathcal{L}} \mathcal{N}_q\,(0, Id)\,.$$

Sous des conditions de régularité, on a alors :

$$\sqrt{n}\,\left[\Delta(\hat{\theta}_n)\mathcal{I}_1^{-1}(\hat{\theta}_n)\Delta(\hat{\theta}_n)^T\right]^{-1/2}\,\left[\varphi(\hat{\theta}_n) - \varphi(\theta)\right] \xrightarrow{\mathcal{L}} \mathcal{N}_q\,(0, Id)$$

ce qui permet de donner des intervalles de confiance asymptotiques pour chaque composante de $\varphi(\theta)$.

Chapitre 8

Tests d'hypothèses

Les tests d'hypothèses sont probablement le concept le plus important de la statistique car c'est ce qui permet de prendre des décisions au vu de résultats d'expériences ou d'observation de phénomènes dans un contexte incertain. Cette notion a été introduite en PS1 sous une forme réduite. Le but de ce chapitre est d'approfondir cette notion. Nous verrons en particulier les notions d'erreur de première et deuxième espèce, de région critique, de puissance et de tests unilatéraux. Nous étudierons le lien entre tests d'hypothèses et intervalles de confiance. Enfin, quelques tests usuels seront présentés.

8.1 Introduction : le problème de décision

Dans tous les domaines, de l'expérimentation scientifique à la vie quotidienne, on est amené à prendre des décisions sur une activité risquée au vu de résultats d'expériences ou d'observation de phénomènes dans un contexte incertain. Par exemple :

- *essais thérapeutiques* : décider si un nouveau traitement médical est meilleur qu'un ancien au vu du résultat de son expérimentation sur des malades.
- *santé* : trancher sur la nocivité ou non des OGM ou des antennes de téléphonie mobile, décider s'il faut vacciner toute une population contre le covid.
- *politique* : au vu des résultats d'un sondage, pronostiquer le résultat d'une élection.
- *sociologie* : déterminer s'il existe une inégalité de genre dans l'accès à l'emploi.
- *finance* : au vu du marché, décider si on doit ou pas se lancer dans une opération financière donnée.
- *justice* : décider si l'accusé est innocent ou coupable à partir des informations acquises pendant le procès.

Dans chaque cas, le **problème de décision** consiste à trancher, au vu d'observations, entre une hypothèse appelée **hypothèse nulle**, notée H_0, et une autre hypothèse dite **hypothèse alternative**, notée H_1. En général, on suppose qu'une et une seule de ces deux hypothèses est vraie. Un **test d'hypothèses** est une procédure qui permet de choisir entre ces deux hypothèses.

Dans un problème de décision, deux types d'erreurs sont possibles :

- **erreur de première espèce** : décider que H_1 est vraie alors que H_0 est vraie.
- **erreur de seconde espèce** : décider que H_0 est vraie alors que H_1 est vraie.

Les conséquences de ces deux erreurs peuvent être d'importances diverses. En général, une des erreurs est plus grave que l'autre :

- *essais thérapeutiques* : on peut adopter un nouveau traitement moins efficace, voire pire que l'ancien, ou se priver d'un nouveau traitement plus efficace que l'ancien.
- *santé* : on peut dépenser des milliards d'euros en vaccins inutiles ou subir une pandémie grave à large échelle.

- *finance* : si on décide à tort que l'on peut lancer l'opération, on risque de perdre beaucoup d'argent ; si on décide à tort de ne pas lancer l'opération, on peut se priver d'un bénéfice important.
- *justice* : on peut condamner un innocent ou acquitter un coupable.

A toute décision correspond une probabilité de décider juste et une probabilité de se tromper :

- La probabilité de l'erreur de première espèce, qui est la probabilité de rejeter à tort H_0, est notée α et est appelée **seuil** ou **niveau de signification** du test. C'est la même terminologie que pour les intervalles de confiance, ce qui n'est pas un hasard, comme nous le verrons plus loin.
- La probabilité de l'erreur de deuxième espèce est notée $1 - \beta$.
- β est la probabilité de décider H_1 (ou de rejeter H_0) à raison. Elle est appelée **puissance** du test.
- $1 - \alpha$ est appelé **niveau de confiance** du test.

Le tableau 8.1 résume simplement le rôle de ces probabilités.

Vérité / Décision	H_0	H_1
H_0	$1-\alpha$	$1-\beta$
H_1	α	β

TABLE 8.1 – probabilités de bonne et mauvaise décision dans un test d'hypothèses

L'idéal serait évidemment de trouver une procédure qui minimise les deux risques d'erreur en même temps. Malheureusement, on montre qu'ils varient en sens inverse, c'est-à-dire que toute procédure diminuant α va en général augmenter $1-\beta$ et réciproquement. Dans la pratique, on va donc considérer que l'une des deux erreurs est plus importante que l'autre, et tâcher d'éviter que cette erreur se produise. Il est alors possible que l'autre erreur survienne. Par exemple, dans le cas du procès, on fait en général tout pour éviter de condamner un innocent, quitte à prendre le risque d'acquitter un coupable.

On va choisir H_0 et H_1 de sorte que l'erreur que l'on cherche à éviter soit l'erreur de première espèce. Mathématiquement cela revient à se fixer la valeur du seuil du test α. Plus la conséquence de l'erreur est grave, plus α sera choisi petit. Les valeurs usuelles de α sont 10%, 5%, 1%, ou beaucoup moins. Le **principe de précaution** consiste à limiter au maximum la probabilité de se tromper, donc à prendre α très petit.

On appelle **règle de décision** une règle qui permette de choisir entre H_0 et H_1 au vu de l'observation x, sous la contrainte que la probabilité de rejeter à tort H_0 est égale à α fixé. Une idée naturelle est de conclure que H_0 est fausse s'il est très peu probable d'observer x quand H_0 est vraie.

Par exemple, admettons que l'on doive décider si une pièce est truquée ou pas au vu de 100 lancers de cette pièce. Si on observe 90 piles, il est logique de conclure que la pièce est truquée et on pense avoir une faible probabilité de se tromper en concluant cela. Mais si on observe 65 piles, que conclure ?

On appelle **région critique** du test, et on note W, l'ensemble des valeurs de l'observation x pour lesquelles on rejettera H_0. La région critique est souvent déterminée à l'aide du bon sens. Sinon, on utilisera une fonction pivotale ou des théorèmes d'optimalité. W dépend du seuil α et est déterminée a priori, indépendamment de la valeur des observations. Ensuite, si les observations appartiennent à W, on rejette H_0, sinon on ne la rejette pas.

Remarque 8.1.1. Il vaut mieux dire "*ne pas rejeter* H_0" que "*accepter* H_0". En effet, si on rejette H_0, c'est que les observations sont telles qu'il est très improbable que H_0 soit vraie. Si on ne rejette pas H_0, c'est qu'on ne dispose pas de critères suffisants pour pouvoir dire que H_0 est fausse. Mais cela ne veut pas dire que H_0 est vraie. Un test permet de dire qu'une hypothèse est très probablement fausse ou seulement peut-être vraie. Par exemple, si on n'a pas de preuve qu'un accusé est coupable, cela ne veut pas forcément dire qu'il est innocent (et réciproquement).

Par conséquent, dans un problème de test, il faut choisir les hypothèses H_0 et H_1 de façon à ce que ce qui soit vraiment intéressant, c'est de rejeter H_0. En pratique, on verra plus tard que la nature des données imposera de fait un choix pertinent des hypothèses.

Récapitulons l'ensemble de la démarche à suivre pour effectuer un test d'hypothèses :

1. Choisir H_0 et H_1 de sorte que ce qui importe, c'est de ne pas se tromper en rejetant H_0.
2. Se fixer α selon la gravité des conséquences de l'erreur de première espèce.
3. Déterminer la région critique W.
4. Regarder si les observations se trouvent ou pas dans W.
5. Prendre la décision, c'est-à-dire conclure au rejet ou au non-rejet de H_0.

Pour le même problème de décision, plusieurs tests (c'est-à-dire plusieurs régions critiques) de même seuil sont souvent possibles. Dans ce cas, le meilleur de ces tests est celui qui minimisera la probabilité de l'erreur de seconde espèce, c'est à dire celui qui maximisera la puissance β. Le meilleur des tests possibles de seuil fixé est le **test le plus puissant**. Il arrive que l'on puisse le déterminer, mais pas toujours.

Dans un modèle statistique quelconque $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, où l'on observe la réalisation x d'un élément aléatoire X de loi de probabilité $\mathbb{P}^X$, les questions que l'on se pose portent sur la loi des observations. Donc on s'intéressera à des tests de :

$$H_0 : \text{“}\mathbb{P}^X \in \mathcal{P}_0\text{”} \text{ contre } H_1 : \text{“}\mathbb{P}^X \in \mathcal{P}_1\text{”}.$$

où $\mathcal{P}_0$ et $\mathcal{P}_1$ sont des familles de lois de probabilité incluses dans $\mathcal{P}$.

Dans ce cours, on se contentera de considérer des modèles paramétriques d'échantillon pour un paramètre θ réel. Dans ce cas, les hypothèses d'un test portent sur la valeur de θ. Les tests de ce type sont appelés **tests paramétriques**. Les tests qui ne portent pas sur la valeur d'un paramètre sont appelés **tests non paramétriques**. Il en existe de tous les types.

8.2 Formalisation du problème de test paramétrique sur un échantillon

On se place dans un modèle paramétrique d'échantillon de taille n $(\mathcal{X}, \mathcal{A}, \{P_X^\theta; \theta \in \Theta \subset \mathbb{R}\})^n$. Les observations sont les réalisations $x_1, \ldots, x_n$ de variables aléatoires $X_1, \ldots, X_n$ indépendantes et de même loi, dépendant d'un paramètre réel inconnu θ. Pour un paramètre de dimension $d > 1$, on fera des tests sur chacune de ses composantes. Par exemple, on fera des tests sur la moyenne de la loi normale, puis des tests sur la variance, mais pas sur les deux en même temps.

Une **hypothèse** est **simple** si elle est du type “$\theta = \theta_0$”, où θ_0 est un réel fixé. Une **hypothèse** est **composite** ou **multiple** si elle est du type “$\theta \in A$” où A est une partie de $\mathbb{R}$ non réduite à un élément.

8.2.1 Tests d'hypothèses simples

Un **test d'hypothèses simples** est un test dans lequel les hypothèses nulle et alternative sont simples toutes les deux. C'est donc un test du type

$$H_0 : \text{“}\theta = \theta_0\text{”} \text{ contre } H_1 : \text{“}\theta = \theta_1\text{”}.$$

Un tel test est un cas d'école : il permet de dire laquelle des deux valeurs θ_0 et θ_1 est la plus vraisemblable au vu des observations. Mais il ne prend pas en compte la possibilité que θ ne soit égal ni à θ_0 ni à θ_1. Pour cela, il faudra faire un test d'hypothèses composites.

Le seuil du test est la probabilité de rejeter à tort H_0, c'est à dire la probabilité que les observations soient dans la région critique W quand la vraie valeur de θ est θ_0 :

$$\alpha = \mathbb{P}\left((X_1, \ldots, X_n) \in W; \theta_0\right)$$

La puissance du test est la probabilité de rejeter à raison H_0, c'est à dire la probabilité que les observations soient dans la région critique quand la vraie valeur de θ est θ_1 :

$$\beta = \mathbb{P}\left((X_1, \ldots, X_n) \in W; \theta_1\right)$$

8.2.2 Tests d'hypothèses composites

Un **test d'hypothèses composites** est un test dans lequel l'une au moins des deux hypothèses est composite. Les tests les plus usuels sont du type :

- **test bilatéral** : H_0 : "$\theta = \theta_0$" contre H_1 : "$\theta \neq \theta_0$" (seule H_1 est composite). C'est ce cas qui a été vu en PS1.

- **tests unilatéraux** : H_0 : "$\theta \leq \theta_0$" contre H_1 : "$\theta > \theta_0$" ou H_0 : "$\theta \geq \theta_0$" contre H_1 : "$\theta < \theta_0$" (H_0 et H_1 sont composites).

On pourrait aussi imaginer des tests du type H_0 : "$\theta \in [\theta_1, \theta_2]$" contre H_1 : "$\theta < \theta_1$ ou $\theta > \theta_2$". Toutes les variantes sont envisageables. Dans tous ces exemples, H_0 et H_1 sont complémentaires : des deux hypothèses, l'une est forcément vraie. C'est ce cas qui est important en pratique.

Quand une hypothèse est composite, la notion de puissance est à repréciser. En effet, β a été définie comme la probabilité de rejeter H_0 à raison, c'est à dire de rejeter H_0 quand H_1 est vraie. Or, dans les exemples ci-dessus, il y a une infinité de valeurs de θ pour lesquelles H_1 est vraie. Donc la puissance du test doit dépendre de la vraie valeur (inconnue) de θ, ce qui nous amène à redéfinir la puissance et le seuil d'un test :

Définition 8.2.1. La **puissance** d'un test portant sur la valeur d'un paramètre réel θ est la fonction de θ définie par :

$$\begin{aligned} \beta : \quad & \mathbb{R} \rightarrow [0,1] \\ & \theta \mapsto \beta(\theta) = \mathbb{P}\left((X_1, \ldots, X_n) \in W ; \theta\right) \end{aligned}$$

Le **seuil** du test est $\alpha = \underset{H_0}{Sup}\, \beta(\theta)$.

$\beta(\theta)$ est la probabilité de rejeter H_0 quand la vraie valeur du paramètre est θ. $\alpha = \underset{H_0}{Sup}\, \beta(\theta)$ est la probabilité maximale de rejeter H_0 alors que H_0 est vraie, c'est à dire la plus forte probabilité de rejeter à tort H_0. Par exemple, pour un test bilatéral, $\alpha = \beta(\theta_0)$, et pour le premier test unilatéral présenté, $\alpha = \underset{\theta \leq \theta_0}{Sup}\, \beta(\theta)$.

Une fois H_0 et H_1 déterminées et α fixé, il faut construire la région critique W. Pour comprendre comment déterminer une région critique, nous allons détailler dans la section suivante la construction d'un test sur la moyenne d'une loi normale, à partir d'un exemple introductif.

8.3 Tests sur la moyenne d'une loi normale

8.3.1 Exemple introductif : essais thérapeutiques

Pour apaiser un certain type de maux de tête, on a l'habitude de traiter les malades avec un médicament A. Une étude statistique a montré que la durée de disparition de la douleur chez les malades traités avec A était une variable aléatoire de loi normale $\mathcal{N}\left(m_0, \sigma_0^2\right)$, avec $m_0 = 30$ mn et $\sigma_0 = 5$ mn. Un laboratoire pharmaceutique a conçu un nouveau médicament B et désire tester son efficacité. Pour cela, le nouveau médicament a été administré à n malades cobayes, et on a mesuré la durée de disparition de la douleur pour chacun d'entre eux : $x_1, \ldots, x_n$. Une étude de statistique descriptive sur ces données a amené les pharmacologues à considérer que cette durée était une variable aléatoire de loi normale $\mathcal{N}\left(m, \sigma^2\right)$.

Remarque 8.3.1. En toute rigueur, on ne devrait pas modéliser une durée (positive) par une variable aléatoire qui, comme pour la loi normale, peut prendre des valeurs négatives. En pratique, on peut le faire quand, pour les lois considérées, la probabilité que la variable soit négative est négligeable.

L'effet du nouveau médicament se traduit facilement sur la valeur de la durée moyenne de disparition de la douleur :

- "$m = m_0$" : le médicament B a en moyenne le même effet que le médicament A.
- "$m < m_0$" : le médicament B est en moyenne plus efficace que le médicament A.
- "$m > m_0$" : le médicament B est en moyenne moins efficace que le médicament A.

Nous reviendrons ultérieurement sur l'interprétation de la valeur de l'écart-type σ en termes d'efficacité du médicament.

Pour savoir s'il faut commercialiser B, il faut trancher entre ces 3 hypothèses. L'important est de ne pas se tromper si on décide de changer de médicament : il est préférable de conserver un médicament moins performant que le nouveau que d'adopter un médicament moins performant que l'ancien. Il faut donc que l'hypothèse "$m < m_0$" corresponde au rejet de H_0.

Par conséquent, nous allons tester H_0 : "$m \geq m_0$" contre H_1 : "$m < m_0$" au vu de n réalisations indépendantes $x_1, \ldots, x_n$ de la loi $\mathcal{N}\left(m, \sigma^2\right)$.

8.3.2 Première idée

Puisque $\bar{X}_n$ est l'ESBVM de m, une première idée est de conclure que $m < m_0$ si et seulement si $\bar{x}_n < m_0$: la durée moyenne de disparition de la douleur sur les malades traités avec B est plus petite que ce qu'elle est sur les malades traités avec A. Cela revient à proposer comme région critique du test :

$$W = \{(x_1, \ldots, x_n); \bar{x}_n < m_0\}$$

Si $\bar{x}_n$ est beaucoup plus petite que m_0, il est en effet très probable que B soit plus efficace que A. Mais si $\bar{x}_n$ est proche de m_0 tout en étant plus petite, on risque de se tromper si on affirme que $m < m_0$. La probabilité de cette erreur, qui n'est autre que le risque de première espèce α, est très facile à calculer :

$$\begin{aligned} \alpha &= \underset{H_0}{Sup}\, \beta(m) = \underset{m \geq m_0}{Sup}\ \mathbb{P}\left(\bar{X}_n < m_0; m\right) \\ &= \underset{m \geq m_0}{Sup}\ \mathbb{P}\left(\sqrt{n}\,\frac{\bar{X}_n - m}{\sigma} < \sqrt{n}\,\frac{m_0 - m}{\sigma}; m\right) = \underset{m \geq m_0}{Sup}\ \phi\left(\sqrt{n}\,\frac{m_0 - m}{\sigma}\right) \end{aligned}$$

où ϕ est la fonction de répartition de la loi normale centrée-réduite. En effet, comme on l'a déjà vu, si $X_1, \ldots, X_n$ sont indépendantes et de même loi $\mathcal{N}\left(m, \sigma^2\right)$, alors $\bar{X}_n$ est de loi $\mathcal{N}\left(m, \frac{\sigma^2}{n}\right)$ et $\sqrt{n}\,\frac{\bar{X}_n - m}{\sigma}$ est de loi $\mathcal{N}(0, 1)$.

$\phi(u)$ est une fonction croissante de u, donc $\beta(m) = \phi\left(\sqrt{n}\,\frac{m_0 - m}{\sigma}\right)$ est une fonction décroissante de m. Par conséquent, $\alpha = \underset{m \geq m_0}{Sup}\ \beta(m) = \beta(m_0) = \phi(0) = 1/2$.

Il y a donc une chance sur deux de se tromper si on décide que B est plus efficace que A quand $\bar{x}_n < m_0$. C'est évidemment beaucoup trop.

8.3.3 Deuxième idée

On voit qu'il faut en fait rejeter H_0 quand $\bar{x}_n$ est *significativement plus petit* que m_0. Cela revient à prendre une région critique de la forme :

$$W = \{(x_1, \ldots, x_n); \bar{x}_n < l_\alpha\}, \text{ où } l_\alpha < m_0.$$

La borne l_α dépend du seuil α que l'on s'est fixé. Moins on veut risquer de rejeter à tort H_0, plus α sera petit, et plus l_α sera petit. Le sens de l'expression *significativement plus petit* est lié à la valeur de α.

Un calcul analogue au précédent montre que :

$$\alpha = \underset{H_0}{Sup}\, \beta(m) = \underset{m \geq m_0}{Sup}\ \mathbb{P}\left(\bar{X}_n < l_\alpha; m\right) = \underset{m \geq m_0}{Sup}\ \phi\left(\sqrt{n}\,\frac{l_\alpha - m}{\sigma}\right) = \phi\left(\sqrt{n}\,\frac{l_\alpha - m_0}{\sigma}\right)$$

On obtient donc $\sqrt{n}\,\frac{l_\alpha - m_0}{\sigma} = \phi^{-1}(\alpha)$, d'où $l_\alpha = m_0 + \frac{\sigma}{\sqrt{n}}\,\phi^{-1}(\alpha) = m_0 - \frac{\sigma}{\sqrt{n}}\,u_{2\alpha}$, avec les notations habituelles pour les quantiles de la loi normale.

En conclusion, on a :

Proposition 8.3.1. *Un test de seuil α de H_0 : "$m \geq m_0$" contre H_1 : "$m < m_0$" est déterminé par la région critique :*

$$W = \left\{(x_1, \ldots, x_n); \bar{x}_n < m_0 - \frac{\sigma}{\sqrt{n}} u_{2\alpha}\right\}$$

8.3.4 Troisième idée

La région critique proposée ci-dessus pose un problème : ce test est inutilisable si on ne connaît pas la vraie valeur de σ, ce qui est toujours le cas en pratique. Pour pallier cet inconvénient, il est logique de remplacer σ par son estimateur S'_n, ce qui nécessite de remplacer la loi normale par la loi de Student.

Rappelons en effet que si $X_1, \ldots, X_n$ sont indépendantes et de même loi $\mathcal{N}\left(m, \sigma^2\right)$, alors $\sqrt{n}\,\frac{\bar{X}_n - m}{S'_n}$ est de loi $St(n-1)$. Alors, à partir d'une région critique de la forme $W = \{(x_1, \ldots, x_n); \bar{x}_n < l_\alpha\}$, on obtient :

$$\begin{aligned}
\alpha &= \underset{H_0}{Sup}\,\beta(m) = \underset{m \geq m_0}{Sup}\ \mathbb{P}\left(\bar{X}_n < l_\alpha; m\right) = \underset{m \geq m_0}{Sup}\ \mathbb{P}\left(\sqrt{n}\,\frac{\bar{X}_n - m}{S'_n} < \sqrt{n}\,\frac{l_\alpha - m}{S'_n}; m\right) \\
&= \underset{m \geq m_0}{Sup}\ F_{St(n-1)}\left(\sqrt{n}\,\frac{l_\alpha - m}{S'_n}\right) = F_{St(n-1)}\left(\sqrt{n}\,\frac{l_\alpha - m_0}{S'_n}\right)
\end{aligned}$$

D'où $\sqrt{n}\,\frac{l_\alpha - m_0}{S'_n} = F^{-1}_{St(n-1)}(\alpha)$. On note $t_{n,\alpha} = F^{-1}_{St(n)}\left(1 - \frac{\alpha}{2}\right)$ le quantile d'ordre $1 - \alpha/2$ de la loi $St(n)$, qu'on peut obtenir en `R` avec la commande `qt(1-alpha/2,n)` ou à l'aide des tables du chapitre 9. Par symétrie de la loi de Student, $F^{-1}_{St(n-1)}(\alpha) = -t_{n-1,2\alpha}$. Finalement, $l_\alpha = m_0 - \frac{S'_n}{\sqrt{n}}\,t_{n-1,2\alpha}$.

En conclusion, on a :

Proposition 8.3.2. *Un test de seuil α de H_0 : "$m \geq m_0$" contre H_1 : "$m < m_0$" est déterminé par la région critique :*

$$W = \left\{(x_1, \ldots, x_n); \bar{x}_n < m_0 - \frac{s'_n}{\sqrt{n}}\,t_{n-1,2\alpha}\right\}$$

Remarque : La région critique peut aussi s'écrire :

$$W = \left\{(x_1, \ldots, x_n); \sqrt{n}\,\frac{\bar{x}_n - m_0}{s'_n} < -t_{n-1,2\alpha}\right\}$$

Cette forme met en évidence l'utilisation de la variable aléatoire $\sqrt{n}\,\frac{\bar{X}_n - m_0}{S'_n}$ qui n'est autre que la fonction pivotale qui est utilisée pour déterminer un intervalle de confiance pour m (voir PS1). C'est cette forme que l'on conservera dans la suite.

8.3.5 Exemple

Avec le médicament A, la durée moyenne de disparition de la douleur était 30 mn. On a administré le médicament B à 12 malades et relevé les durées de disparition de la douleur suivants :

25 28 20 32 17 24 41 28 25 30 27 24

La moyenne empirique de ces données est $\bar{x}_n$= 26.75 et l'écart-type estimé est $s'_n = 6.08$.

On décide de ne commercialiser B que si on est sûr à 95% qu'il est plus efficace que A. Cela revient donc à faire un test de H_0 : "$m \geq 30$" contre H_1 : "$m < 30$" au seuil $\alpha = 5\%$.

On voit qu'il s'agit finalement de déterminer si 26.75 est suffisamment inférieur à 30 pour que l'on puisse conclure que le médicament B réduit vraiment la durée de disparition de la douleur.

D'après ce qui précède, on rejettera H_0 si $\sqrt{n}\,\frac{\bar{x}_n - m_0}{s'_n} < -t_{n-1,2\alpha}$.

Or $\sqrt{n}\,\frac{\bar{x}_n - m_0}{s'_n} = \sqrt{12}\,\frac{26.75 - 30}{6.08} = -1.853$ et $t_{n-1,2\alpha} = t_{11,0.1} = 1.796$.

$-1.853 < -1.796$, donc les observations sont dans la région critique. On rejette donc H_0, ce qui signifie que l'on conclut que B est plus efficace que A, avec moins de 5% de chances de se tromper. Par conséquent, on peut lancer la commercialisation du médicament B.

8.3.6 La p-valeur

On voit ici le rôle fondamental du seuil α. Si on avait pris $\alpha = 1\%$, on aurait eu $t_{11,0.02} = 2.718$. Comme $-1.853 > -2.718$, on n'aurait pas rejeté H_0, donc on n'aurait pas adopté le médicament B.

Ce phénomène est normal : se fixer un seuil α petit revient à éviter au maximum d'adopter à tort le médicament B. Or un bon moyen de ne pas prendre ce risque, c'est de conserver le médicament A. Le test de seuil α = 0 consiste à conserver le médicament A quelles que soient les observations : la probabilité de rejeter à tort H_0 est nulle quand on ne rejette jamais H_0 ! En pratique, plus α est petit, moins on aura tendance à rejeter H_0. D'une certaine façon, cela signifie que le principe de précaution conduit au conservatisme...

Il est donc fondamental de bien savoir évaluer les risques et de choisir α en connaissance de cause.

Cet exemple avec $\alpha = 1\%$ permet également de comprendre la nuance entre "ne pas rejeter H_0" et "accepter H_0" : on va conclure que rien ne prouve que B est plus efficace que A, mais on ne va évidemment pas conclure que A est plus efficace que B.

La remarque précédente met en évidence l'existence d'un seuil critique α_c tel que :
- pour tout seuil $\alpha > \alpha_c$, on rejettera H_0,
- pour tout seuil $\alpha \leq \alpha_c$, on ne rejettera pas H_0.

C'est donc la valeur α_c de α pour laquelle on aura une égalité dans la région critique. Cette valeur est appelée la **p-valeur**.

Dans notre exemple, α_c vérifie $\sqrt{n}\,\dfrac{\bar{x}_n - m_0}{s'_n} = -t_{n-1,2\alpha_c}$. La table de la loi de Student permet de constater que $-t_{11,0.05} = -2.201 < -1.853 < -1.796 = -t_{11,0.1}$. On en déduit que $5\% < 2\alpha_c < 10\%$, d'où $2.5\% < \alpha_c < 5\%$.

Pour calculer exactement la p-valeur, on écrit :

$$-t_{n-1,2\alpha_c} = F^{-1}_{St(n-1)}\left(\frac{2\alpha_c}{2}\right) = F^{-1}_{St(n-1)}(\alpha_c) \implies \alpha_c = F_{St(n-1)}\left(\sqrt{n}\,\frac{\bar{x}_n - m_0}{s'_n}\right)$$

et on obtient ici $\alpha_c = 0.04547$.

La réponse à un test de seuil fixé est binaire : on rejette ou on ne rejette pas H_0. Fournir une p-valeur est une réponse plus riche puisqu'elle permet de connaître le résultat du test pour n'importe quel choix du seuil. C'est pourquoi le traitement des tests d'hypothèses par les logiciels de statistique consiste à fournir des p-valeurs.

En `R`, la commande permettant d'effectuer un test sur la moyenne d'une loi normale est `t.test`. L'option `alternative` permet de préciser lequel du test bilatéral et des deux tests unilatéraux on choisit. Sur l'exemple, on obtient :

```
>  medic<-c(25,28,20,32,17,24,41,28,25,30,27,24)
>  t.test(medic,alternative="less",mu=30)

        One Sample t-test

data: medic
t = -1.8526, df = 11, p-value = 0.04547
alternative hypothesis: true mean is less than 30
95 percent confidence interval:
      -Inf 29.90056
sample estimates:
mean of x
     26.75
```

La p-valeur est ici $\alpha_c = 4.5\%$. Cela signifie que, pour tout seuil supérieur à 4.5% (c'est le cas de 5%), on rejettera H_0, donc on conclura que B est plus efficace que A, et pour tout seuil inférieur à 4.5% (c'est le cas de 1%), on ne rejettera pas H_0, donc on conclura que B n'est pas plus efficace que A.

De manière générale, la p-valeur peut être comprise comme étant la probabilité, sous l'hypothèse nulle, que la statistique de test soit encore plus extrême que ce qui a été observé. Si cette probabilité est forte, il n'y a pas de raison de douter de la véracité de H_0. Mais si elle est faible, on peut en douter. Donc plus la p-valeur est petite, moins on prend de risque en rejetant H_0.

8.3.7 Remarques

Remarque 8.3.2. Pour des raisons de symétrie, un test de "$m \leq m_0$" contre "$m > m_0$" aura pour région critique

$$W = \left\{(x_1, \ldots, x_n); \sqrt{n}\,\frac{\bar{x}_n - m_0}{s'_n} > t_{n-1,2\alpha}\right\}$$

Remarque 8.3.3. Dans l'exemple des essais thérapeutiques, $\sqrt{n}\,\dfrac{\bar{x}_n - m_0}{s'_n} = -1.853$. $\forall \alpha$, $t_{n-1,2\alpha} > 0$, donc on n'aura jamais $\sqrt{n}\,\dfrac{\bar{x}_n - m_0}{s'_n} > t_{n-1,2\alpha}$. Par conséquent, quel que soit α, on ne rejettera pas H_0, donc on ne conclura pas que $m > m_0$. Cela signifie que, pour ces données, faire ce test n'a pas de sens.

En effet, $m_0 = 30$ mn et $\bar{x}_n = 26.75$ mn. En observant une moyenne empirique inférieure à 30, on ne peut pas en conclure que la vraie moyenne m est supérieure à 30, quel que soit le risque que l'on prend. La seule question valable est de déterminer si le fait que $\bar{x}_n = 26.75$ peut permettre de conclure que $m \leq 30$ avec un bon degré de confiance. On voit donc que si, en théorie, on peut choisir de tester "$m \geq m_0$" contre "$m < m_0$" ou bien de tester "$m \leq m_0$" contre "$m > m_0$", en pratique, la nature des données fera qu'un seul de ces deux tests aura du sens.

Remarque 8.3.4. Pour le test bilatéral de H_0 : "$m = m_0$" contre H_1 : "$m \neq m_0$", le bon sens veut que l'on rejette H_0 si $\bar{x}_n$ est significativement éloigné de m_0. On prendra donc une région critique du type $W = \{(x_1, \ldots, x_n); |\bar{x}_n - m_0| > l_\alpha\}$. Alors, comme précédemment on obtient :

$$\begin{aligned}
\alpha &= \underset{m=m_0}{Sup}\ \mathbb{P}\left(\left|\bar{X}_n - m_0\right| > l_\alpha; m\right) = \mathbb{P}\left(\left|\bar{X}_n - m_0\right| > l_\alpha; m_0\right) \\
&= \mathbb{P}\left(\left|\sqrt{n}\,\frac{\bar{X}_n - m_0}{S'_n}\right| > \sqrt{n}\,\frac{l_\alpha}{S'_n}; m_0\right)
\end{aligned}$$

On en déduit que $\sqrt{n}\,\dfrac{l_\alpha}{S'_n} = t_{n-1,\alpha}$, d'où $l_\alpha = \dfrac{S'_n}{\sqrt{n}}\,t_{n-1,\alpha}$. On obtient donc comme région critique :

$$W = \left\{(x_1, \ldots, x_n); |\bar{x}_n - m_0| > \frac{s'_n}{\sqrt{n}}\,t_{n-1,\alpha}\right\} = \left\{(x_1, \ldots, x_n); \left|\sqrt{n}\,\frac{\bar{x}_n - m_0}{s'_n}\right| > t_{n-1,\alpha}\right\}$$

C'est ce test bilatéral qui a été vu en PS1. Posons $t = \sqrt{n}\,\dfrac{\bar{x}_n - m_0}{s'_n}$ la valeur observée de la statistique de test. La p-valeur α_c du test vérifie $|t| = t_{n-1,\alpha_c} = F^{-1}_{St(n-1)}(1 - \frac{\alpha_c}{2})$. On en déduit que $1 - \frac{\alpha_c}{2} = F_{St(n-1)}(|t|) = \mathbb{P}(T_n \leq |t|)$, où T_n est une variable aléatoire de loi $St(n-1)$. D'où $\alpha_c = 2\mathbb{P}(T_n > |t|)$ et on retrouve bien le résultat vu en PS1.

Remarque 8.3.5. Pour alléger les écritures, on écrit souvent une région critique en omettant l'expression $(x_1, \ldots, x_n)$;, ce qui donne par exemple $W = \left\{\sqrt{n}\,\dfrac{\bar{x}_n - m_0}{s'_n} < -t_{n-1,2\alpha}\right\}$. Mais il faut toujours garder à l'esprit que la région critique est l'ensemble des valeurs des observations pour lesquelles on rejettera H_0.

Remarque 8.3.6. En PS1, on a donné le résultat d'un test d'hypothèse en comparant la p-valeur à 5%. Cela signifie qu'on a systématiquement effectué des tests de seuil 5%. La démarche présentée ici permet de considérer n'importe quel seuil, ce qui permet de moduler la décision issue du test en fonction de la gravité de ses conséquences.

8.3.8 Les tests de Student

Finalement, on dispose d'une procédure permettant d'effectuer le test bilatéral et les deux tests unilatéraux portant sur la moyenne de la loi normale. Ces trois tests sont connus sous le nom de **tests de Student**.

Proposition 8.3.3. *Tests de Student sur la moyenne d'une loi normale.*

- *Test de "$m \leq m_0$" contre "$m > m_0$" :* $W = \left\{\sqrt{n}\,\dfrac{\bar{x}_n - m_0}{s'_n} > t_{n-1,2\alpha}\right\}$.
- *Test de "$m \geq m_0$" contre "$m < m_0$" :* $W = \left\{\sqrt{n}\,\dfrac{\bar{x}_n - m_0}{s'_n} < -t_{n-1,2\alpha}\right\}$.
- *Test de "$m = m_0$" contre "$m \neq m_0$" :* $W = \left\{\left|\sqrt{n}\,\dfrac{\bar{x}_n - m_0}{s'_n}\right| > t_{n-1,\alpha}\right\}$.

Remarque 8.3.7. Les tests ci-dessus ont été présentés comme des tests portant sur la valeur de la moyenne d'une loi normale. En fait, grâce au théorème central-limite, on sait que, quand n est assez grand, $\bar{X}_n$ est approximativement de loi normale, quelle que soit la loi de probabilité des observations. Cette propriété permet de montrer que pour n suffisamment grand (en pratique $n \geq 30$), on pourra utiliser le test de Student pour faire un test sur la valeur de la moyenne de n'importe quelle loi de probabilité. On dit que le test de Student est **robuste** à la non-normalité.

8.4 Lien entre tests d'hypothèses et intervalles de confiance

Dans le test bilatéral, on rejette l'hypothèse "$m = m_0$" à condition que $\left|\sqrt{n}\,\dfrac{\bar{x}_n - m_0}{s'_n}\right| > t_{n-1,\alpha}$. Or :

$$\begin{aligned}
\left|\sqrt{n}\,\frac{\bar{x}_n - m_0}{s'_n}\right| > t_{n-1,\alpha} \quad &\Leftrightarrow \quad \bar{x}_n - m_0 < -\frac{s'_n}{\sqrt{n}}\,t_{n-1,\alpha} \text{ ou } \bar{x}_n - m_0 > +\frac{s'_n}{\sqrt{n}}\,t_{n-1,\alpha} \\
&\Leftrightarrow \quad m_0 < \bar{x}_n - \frac{s'_n}{\sqrt{n}}\,t_{n-1,\alpha} \text{ ou } m_0 > \bar{x}_n + \frac{s'_n}{\sqrt{n}}\,t_{n-1,\alpha} \\
&\Leftrightarrow \quad m_0 \notin \left[\bar{x}_n - \frac{s'_n}{\sqrt{n}}\,t_{n-1,\alpha}, \bar{x}_n + \frac{s'_n}{\sqrt{n}}\,t_{n-1,\alpha}\right]
\end{aligned}$$

Cet intervalle n'est autre que l'intervalle de confiance usuel de seuil α pour la moyenne de la loi normale, vu en PS1. On rejette donc "$m = m_0$" si m_0 n'appartient pas à cet intervalle. Il y a donc un lien étroit entre les tests d'hypothèses et les intervalles de confiance.

C'est logique : on a une confiance $1 - \alpha$ dans le fait que m appartient à l'intervalle de confiance. Si m_0 n'appartient pas à cet intervalle, il est vraiment douteux que $m = m_0$. On a même une confiance $1 - \alpha$ dans le fait que $m \neq m_0$. On peut donc construire un test d'hypothèses sur la valeur d'un paramètre à partir d'un intervalle de confiance pour ce paramètre.

Or, pour construire un tel intervalle, on a eu besoin d'une fonction pivotale. Par conséquent, pour construire un test paramétrique, il suffit de connaître une fonction pivotale. Dans le cas de la moyenne de la loi normale, la fonction pivotale est $\sqrt{n}\,\dfrac{\bar{X}_n - m}{S'_n}$.

La dualité entre intervalles de confiance et tests d'hypothèses fait que, en `R`, la commande `t.test` permet à la fois d'effectuer un test et d'obtenir un intervalle de confiance sur la moyenne de la loi normale.

Dans l'exemple, la commande `t.test(medic,mu=30)` (ou, de manière équivalente `t.test(medic, alternative="two.sided",mu=30)`) permet à la fois de faire un test de "$m = 30$" contre "$m \neq 30$", et de donner un intervalle de confiance (bilatéral) pour m au seuil 5% :

```
>  t.test(medic,mu=30)

        One Sample t-test

data: medic
t = -1.8526, df = 11, p-value = 0.09094
alternative hypothesis: true mean is not equal to 30
95 percent confidence interval:  22.88876 30.61124
sample estimates:
mean of x
    26.75
```

L'intervalle de confiance de seuil 5% pour m est $[22.88876, 30.61124]$. 30 est dedans, donc on ne rejette pas l'hypothèse $m = 30$ au seuil 5%. C'est cohérent avec le fait que la p-valeur du test est $9.094\% > 5\%$.

Mais le test bilatéral est ici de peu d'intérêt : ce qui compte, ce n'est pas de savoir si 30 n'est pas une valeur plausible pour m, c'est de déterminer si B est plus efficace que A, c'est-à-dire si $m < 30$. L'application de `t.test` vue plus haut pour effectuer le test unilatéral de "$m \geq 30$" contre "$m < 30$" permet d'obtenir un autre intervalle de confiance (unilatéral) de seuil 5% pour m : $]-\infty, 29.90056]$. 30 n'est pas dans cet intervalle, donc on rejette H_0 et on conclut que $m < 30$ (B est plus efficace que A) au seuil 5%. C'est cohérent avec le fait que la p-valeur du test est $4.547\% < 5\%$.

8.5 Procédure pour construire un test d'hypothèses

Finalement, le plus simple pour construire un test d'hypothèses portant sur la valeur d'un paramètre θ est de se fier à son bon sens. Si on connaît un estimateur $\hat{\theta}_n$ de θ, on procèdera de la façon suivante :

- Test de "$\theta \leq \theta_0$" contre "$\theta > \theta_0$" : on rejette H_0 si $\hat{\theta}_n$ est "trop grand". La région critique est donc de la forme :
$$W = \left\{\hat{\theta}_n > l_\alpha\right\}$$
- Test de "$\theta \geq \theta_0$" contre "$\theta < \theta_0$" : on rejette H_0 si $\hat{\theta}_n$ est "trop petit". La région critique est donc de la forme :
$$W = \left\{\hat{\theta}_n < l_\alpha\right\}$$
- Test de "$\theta = \theta_0$" contre "$\theta \neq \theta_0$" : on rejette H_0 si $\left|\hat{\theta}_n - \theta_0\right|$ est "trop grand" ou bien si $\hat{\theta}_n$ est "soit trop grand, soit trop petit". La région critique est donc de la forme :
$$W = \left\{\hat{\theta}_n < l_{1,\alpha} \text{ ou } \hat{\theta}_n > l_{2,\alpha}\right\}, \text{ avec } l_{1,\alpha} < l_{2,\alpha}$$

Pour déterminer $l_\alpha, l_{1,\alpha}, l_{2,\alpha}$, il faut écrire $\alpha = \underset{H_0}{Sup}\, \mathbb{P}\left((X_1, \ldots, X_n) \in W; \theta\right)$. Par exemple, dans le premier cas, $\alpha = \underset{\theta \leq \theta_0}{Sup}\, \mathbb{P}\left(\hat{\theta}_n > l_\alpha\right)$. Pour pouvoir calculer $\mathbb{P}\left(\hat{\theta}_n > l_\alpha\right)$, il faut utiliser une fonction pivotale.

Malheureusement, cette procédure de bon sens ne permet pas toujours de résoudre le problème. C'est le cas par exemple quand la loi de probabilité de $\hat{\theta}_n$ sous H_0 est complexe et qu'on ne peut pas trouver de fonction pivotale. D'autre part, le test obtenu par cette approche n'est pas forcément optimal, au sens où il peut en exister de plus puissants. Il existe en fait des méthodes statistiques sophistiquées permettant de répondre à ces deux problèmes. Le résultat le plus important est le théorème de Neyman-Pearson.

Le principe "non-rejet $\neq$ acceptation" est à comprendre différemment pour les tests unilatéraux et bilatéraux. Pour les tests unilatéraux, la différence est franche : "B n'est pas plus efficace que A" est différent de "B est moins efficace que A". Pour les tests bilatéraux, H_0 est une hypothèse simple, donc accepter H_0 revient à choisir le modèle correspondant ($\theta = \theta_0$) pour le phénomène étudié. Or tous les modèles sont faux, car ce ne sont que des approximations de la réalité. Ne pas rejeter H_0 consiste à considérer que le modèle correspondant n'est pas absurde. Donc on peut l'adopter, ce qui revient en quelque sorte à "accepter" H_0, au sens où le modèle sous-jacent n'est pas trop mauvais.

8.6 Tests d'hypothèses asymptotiques

Pour $\theta \in \mathbb{R}$, on a vu dans la proposition 7.2.1 qu'un intervalle de confiance asymptotique de seuil α pour θ était :

$$\left[\hat{\theta}_n - \frac{u_\alpha}{\sqrt{n\mathcal{I}_1(\hat{\theta}_n)}}, \hat{\theta}_n + \frac{u_\alpha}{\sqrt{n\mathcal{I}_1(\hat{\theta}_n)}}\right].$$

En adoptant le principe vu dans la section 8.4, on peut construire un test de H_0 : "$\theta = \theta_0$" contre H_1 : "$\theta \neq \theta_0$" consistant à rejeter H_0 si θ_0 n'appartient pas à cet intervalle.

Or $\theta_0 < \hat{\theta}_n - \frac{u_\alpha}{\sqrt{n\mathcal{I}_1(\hat{\theta}_n)}}$ ou $\theta_0 > \hat{\theta}_n + \frac{u_\alpha}{\sqrt{n\mathcal{I}_1(\hat{\theta}_n)}}$ est équivalent à $\sqrt{n\mathcal{I}_1(\hat{\theta}_n)}|\hat{\theta}_n - \theta_0| > u_\alpha$. Donc le test en question a pour région critique $W = \left\{\sqrt{n\mathcal{I}_1(\hat{\theta}_n)}|\hat{\theta}_n - \theta_0| > u_\alpha\right\}$. Le seuil de ce test est

$$\underset{\theta=\theta_0}{Sup}\, \mathbb{P}\left(\sqrt{n\mathcal{I}_1(\hat{\theta}_n)}|\hat{\theta}_n - \theta_0| > u_\alpha\right) = \mathbb{P}_{\theta=\theta_0}\left(\sqrt{n\mathcal{I}_1(\hat{\theta}_n)}|\hat{\theta}_n - \theta_0| > u_\alpha\right)$$

Or on a vu dans la section 7.2.1 que $\sqrt{n\mathcal{I}_1(\hat{\theta}_n)}|\hat{\theta}_n - \theta_0|$ converge en loi vers la loi $\mathcal{N}(0,1)$. Si Y est de loi $\mathcal{N}(0,1)$, $\mathbb{P}(|Y| > u_\alpha) = 1 - \alpha$. Donc :

$$\lim_{n\to+\infty} \mathbb{P}_{\theta=\theta_0}\left(\sqrt{n\mathcal{I}_1(\hat{\theta}_n)}|\hat{\theta}_n - \theta_0| > u_\alpha\right) = \mathbb{P}(|Y| > u_\alpha) = 1 - \alpha$$

Par conséquent, le seuil de ce test tend vers α quand n tend vers l'infini. On dit que c'est un **test asymptotique** de seuil α. Finalement :

Proposition 8.6.1. *Le test d'hypothèses dont la région critique est*

$$W = \left\{\sqrt{n\mathcal{I}_1(\hat{\theta}_n)}|\hat{\theta}_n - \theta_0| > u_\alpha\right\}$$

est un test asymptotique de seuil α de H_0 : "$\theta = \theta_0$" contre H_1 : "$\theta \neq \theta_0$".

De manière analogue, on obtient des tests asymptotiques unilatéraux.

Proposition 8.6.2. *Test asymptotique de seuil α de H_0 : "$\theta \leq \theta_0$" contre H_1 : "$\theta > \theta_0$" :*

$$W = \left\{\sqrt{n\mathcal{I}_1(\hat{\theta}_n)}(\hat{\theta}_n - \theta_0) > u_{2\alpha}\right\}$$

Test asymptotique de seuil α de H_0 : "$\theta \geq \theta_0$" contre H_1 : "$\theta < \theta_0$" :

$$W = \left\{\sqrt{n\mathcal{I}_1(\hat{\theta}_n)}(\hat{\theta}_n - \theta_0) < -u_{2\alpha}\right\}$$

Ces tests asymptotiques sont utiles quand on ne parvient pas à trouver de test exact.

8.7 Tests sur la variance d'une loi normale

On reste dans le modèle d'échantillon de taille n de loi normale $\mathcal{N}(m, \sigma^2)$, mais on s'intéresse cette fois à des tests portant sur la variance σ^2. Par exemple, on souhaite tester H_0 : "$\sigma^2 \leq \sigma_0^2$" contre H_1 : "$\sigma^2 > \sigma_0^2$".

En suivant la démarche présentée ci-dessus, puisque l'ESBVM de σ^2 est ${S'_n}^2$, il est naturel de rejeter H_0 si ${S'_n}^2$ est "trop grand", donc de considérer une région critique de la forme $W = \left\{{s'_n}^2 > l_\alpha\right\}$. Pour calculer $\alpha = \underset{H_0}{Sup}\,\mathbb{P}\left({S'_n}^2 > l_\alpha\right)$, on utilise la fonction pivotale $\dfrac{(n-1){S'_n}^2}{\sigma^2}$, qui est de loi χ^2_{n-1}. On obtient :

$$\begin{aligned}\alpha &= \underset{\sigma^2 \leq \sigma_0^2}{Sup}\ \mathbb{P}({S'_n}^2 > l_\alpha) = \underset{\sigma^2 \leq \sigma_0^2}{Sup}\ \mathbb{P}\left(\frac{(n-1){S'_n}^2}{\sigma^2} > \frac{(n-1)l_\alpha}{\sigma^2}\right) \\ &= \underset{\sigma^2 \leq \sigma_0^2}{Sup}\left[1 - F_{\chi^2_{n-1}}\left(\frac{(n-1)l_\alpha}{\sigma^2}\right)\right] = 1 - F_{\chi^2_{n-1}}\left(\frac{(n-1)l_\alpha}{\sigma_0^2}\right)\end{aligned}$$

D'où $l_\alpha = \dfrac{\sigma_0^2}{n-1} F^{-1}_{\chi^2_{n-1}}(1-\alpha)$. On note $z_{n,\alpha} = F^{-1}_{\chi^2_n}(1-\alpha)$ le quantile d'ordre $1-\alpha$ de la loi χ^2_n, qu'on peut obtenir en `R` avec la commande `qchisq(1-alpha,n)` ou à l'aide des tables du chapitre 9. On a donc $l_\alpha = \dfrac{\sigma_0^2}{n-1} z_{n-1,\alpha}$, et la région critique du test est :

$$W = \left\{{s'_n}^2 > \frac{\sigma_0^2}{n-1} z_{n-1,\alpha}\right\} = \left\{\frac{(n-1){s'_n}^2}{\sigma_0^2} > z_{n-1,\alpha}\right\}$$

On aboutirait au même résultat en partant d'un intervalle de confiance de seuil α pour σ^2 du type $[a, +\infty[$.

Finalement, on obtient :

Proposition 8.7.1. *Tests sur la variance d'une loi normale :*

- *Test de "$\sigma^2 \leq \sigma_0^2$" contre "$\sigma^2 > \sigma_0^2$" :* $W = \left\{\dfrac{(n-1){s'_n}^2}{\sigma_0^2} > z_{n-1,\alpha}\right\}$.
- *Test de "$\sigma^2 \geq \sigma_0^2$" contre "$\sigma^2 < \sigma_0^2$" :* $W = \left\{\dfrac{(n-1){s'_n}^2}{\sigma_0^2} < z_{n-1,1-\alpha}\right\}$.
- *Test de "$\sigma^2 = \sigma_0^2$" contre "$\sigma^2 \neq \sigma_0^2$" :*

$$W = \left\{\frac{(n-1){s'_n}^2}{\sigma_0^2} < z_{n-1,1-\alpha/2}\ ou\ \frac{(n-1){s'_n}^2}{\sigma_0^2} > z_{n-1,\alpha/2}\right\}$$

Dans l'exemple de l'essai thérapeutique, la variance mesure la variabilité de l'effet du médicament. La variabilité est faible si l'effet du médicament est à peu près le même pour tout le monde, et elle est forte si les effets peuvent être très différents d'un individu à un autre. On a évidemment intérêt à avoir une variabilité assez faible pour bien contrôler les effets d'un traitement. Cette variabilité se traduit sur la variance de la loi normale qui modélise la durée de disparition de la douleur chez les malades traités.

Avec le médicament A, l'écart-type était $\sigma_0 = 5$ mn, ce qui signifie que, pour 95% des malades, la douleur disparaît entre $m_0 - 2\sigma_0 = 20$ mn et $m_0 + 2\sigma_0 = 40$ mn. Avec le médicament B, on estime σ par s'_n = 6.08 mn. La variabilité du second médicament est-elle significativement supérieure à celle du premier ?

C'est un test de "$\sigma \leq 5$" contre "$\sigma > 5$" , évidemment identique au test de "$\sigma^2 \leq 25$" contre "$\sigma^2 > 25$". La région critique est $W = \left\{\dfrac{(n-1){s'_n}^2}{\sigma_0^2} > z_{n-1,\alpha}\right\}$.

Au seuil α = 5%, on a $z_{11,5\%} = 19.68$. Et $\frac{(n-1){s'_n}^2}{\sigma_0^2} = \frac{11 \times 6.08^2}{25} = 16.25$.

Comme $16.25 < 19.68$, on n'est pas dans la région critique, donc on ne rejette pas H_0 : on n'a pas de preuves suffisantes pour conclure que la variabilité de l'effet de B est supérieure à celle de A. La différence entre 6.08 et 5 n'est pas significative au seuil choisi.

La p-valeur est obtenue en écrivant :

$$z_{n-1,\alpha_c} = F^{-1}_{\chi^2_{n-1}}(1-\alpha_c) = 16.25 \implies \alpha_c = 1 - F_{\chi^2_{11}}(16.25) = 13.2\%.$$

Donc même au seuil 10%, on ne rejettera pas H_0.

8.8 Test du rapport des vraisemblances maximales

On se place dans un modèle paramétrique $(\mathcal{X}, \mathcal{A}, \{\mathbb{P}_\theta; \theta \in \Theta\})$ et on souhaite tester H_0 : "$\theta \in \Theta_0$" contre H_1 : "$\theta \notin \Theta_0$", où Θ_0 est une partie de Θ. On peut construire un test asymptotique basé sur le rapport de deux fonctions de vraisemblance.

Définition 8.8.1. La **statistique du rapport des vraisemblances maximales** est :

$$v(x) = \frac{\sup\limits_{\theta \in \Theta_0} \mathcal{L}(\theta; x)}{\sup\limits_{\theta \in \Theta} \mathcal{L}(\theta; x)}.$$

Il est clair que $v(x) \in [0,1]$. S'il existe une statistique de maximum de vraisemblance $\hat{\theta}(x)$, le dénominateur est $\sup\limits_{\theta \in \Theta} \mathcal{L}(\theta; x) = \mathcal{L}(\hat{\theta}(x); x)$. Ce dénominateur est la vraisemblance maximale globale alors que le numérateur peut être considéré comme la vraisemblance maximale sous H_0, d'où le nom de statistique du rapport des vraisemblances maximales.

Si $\hat{\theta}(x) \in \Theta_0$, $v(x) = 1$. Comme $\hat{\theta}(x)$ est une bonne estimation de θ, si H_0 est vraie, $v(x)$ ne doit pas être trop loin de 1. Inversement, si $v(x)$ est trop loin de 1, $\hat{\theta}(x) \notin \Theta_0$, donc on peut douter du fait que $\theta \in \Theta_0$. D'où l'idée de construire un test qui va rejeter H_0 si $v(x)$ est trop petit.

Définition 8.8.2. Le **test du rapport des vraisemblances maximales** est le test dont la région critique est de la forme :

$$W = \{v(x) < l_\alpha\}.$$

Autrement dit, on conclura que $\theta \notin \Theta_0$ si et seulement si $v(x) < l_\alpha$. l_α est choisi de sorte que la probabilité maximale de se tromper en rejetant H_0 soit égale à α :

$$\alpha = \sup_{H_0} \mathbb{P}(X \in W) = \sup_{\theta \in \Theta_0} \mathbb{P}(v(X) < l_\alpha).$$

Pour déterminer l_α, il faut connaître la loi de $v(X)$ sous H_0. Donnons le résultat dans un cas particulier.

Proposition 8.8.1. *On considère un modèle d'échantillon $(\mathcal{X}, \mathcal{A}, \{P_\theta; \theta \in \Theta \subset \mathbb{R}^d\})^n$ et le test bilatéral de H_0 : "$\theta = \theta_0$" contre H_1 : "$\theta \neq \theta_0$". On a :*

$$v(x) = \frac{\mathcal{L}(\theta_0; x)}{\sup\limits_{\theta \in \Theta} \mathcal{L}(\theta; x)} = \frac{\mathcal{L}(\theta_0; x)}{\mathcal{L}(\hat{\theta}_n; x)}.$$

Alors, sous H_0, on a :

$$-2 \ln v(X) \xrightarrow{\mathcal{L}} \chi^2_d.$$

Donc le test dont la région critique est

$$W = \{-2 \ln v(x) > z_{d,\alpha}\}$$

où $z_{d,\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi χ^2_d, est asymptotiquement de seuil α pour tester H_0 contre H_1.

Démonstration. On considère le cas où $d = 1$ ($\theta \in \mathbb{R}$) et la loi des observations est continue, de densité f. On utilise un développement limité analogue à celui utilisé pour démontrer les propriétés asymptotiques de l'estimateur de maximum de vraisemblance, en le prenant cette fois à l'ordre 2. Sous H_0 :

$$\ln \mathcal{L}(\theta_0; x) = \ln \mathcal{L}(\hat{\theta}_n; x) + (\theta_0 - \hat{\theta}_n) \frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta; x)_{|\hat{\theta}_n} + \frac{1}{2}(\theta_0 - \hat{\theta}_n)^2 \frac{\partial^2}{\partial \theta^2} \ln \mathcal{L}(\theta; x)_{|\theta'_n},$$

où θ'_n est compris entre θ_0 et $\hat{\theta}_n$.

Par définition de l'EMV, $\frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta; x)_{|\hat{\theta}_n} = 0$. Donc on a :

$$\begin{aligned}
-2 \ln v(X) &= -2 \left[\ln \mathcal{L}(\theta_0; X) - \ln \mathcal{L}(\hat{\theta}_n; X)\right] \\
&= -(\theta_0 - \hat{\theta}_n)^2 \frac{\partial^2}{\partial \theta^2} \ln \mathcal{L}(\theta; X)_{|\theta'_n} \\
&= -(\theta_0 - \hat{\theta}_n)^2 \frac{\partial^2}{\partial \theta^2} \sum_{i=1}^{n} \ln f(X_i; \theta)_{|\theta'_n} \\
&= -\left[\sqrt{n}\,(\theta_0 - \hat{\theta}_n)\right]^2 \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \ln f(X_i; \theta)_{|\theta'_n}
\end{aligned}$$

$\hat{\theta}_n \overset{p.s.}{\rightarrow} \theta_0$ donc $\theta'_n \overset{p.s.}{\rightarrow} \theta_0$. Par la loi des grands nombres :

$$-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \ln f(X_i; \theta)_{|\theta'_n} \overset{p.s.}{\rightarrow} -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \ln f(X_1; \theta)\right]_{|\theta_0} = \mathcal{I}_1(\theta_0).$$

Par ailleurs, $\sqrt{n}\,(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{\mathcal{I}_1(\theta_0)}\right)$, donc $\sqrt{\mathcal{I}_1(\theta_0)}\sqrt{n}\,(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$

et $\mathcal{I}_1(\theta_0) \left[\sqrt{n}\,(\hat{\theta}_n - \theta_0)\right]^2 \xrightarrow{\mathcal{L}} \chi^2_1$, ce qui prouve que $-2 \ln v(X) \xrightarrow{\mathcal{L}} \chi^2_1$.

Revenons à d quelconque : $-2 \ln v(X) \xrightarrow{\mathcal{L}} \chi^2_d$. Le test du rapport des vraisemblances maximales a pour région critique :

$$W = \{v(x) < l_\alpha\} = \{-2 \ln v(x) > -2 \ln l_\alpha\}.$$

Comme H_0 est une hypothèse simple, on cherche donc l_α tel que

$$\alpha = \sup_{H_0} \mathbb{P}(X \in W) = \mathbb{P}_{H_0}(v(X) < l_\alpha) = \mathbb{P}_{H_0}(-2 \ln v(X) > -2 \ln l_\alpha)$$

Comme on ne connaît pas la loi exacte de $-2 \ln v(X)$ sous H_0, on ne peut pas calculer exactement l_α. Mais grâce au résultat asymptotique, on a :

$$\lim_{n \to +\infty} \mathbb{P}_{H_0}(-2 \ln v(X) > -2 \ln l_\alpha) = 1 - F_{\chi^2_d}(-2 \ln l_\alpha).$$

Donc on peut approcher $1 - F_{\chi^2_d}(-2 \ln l_\alpha)$ par α et $-2 \ln l_\alpha$ par $F^{-1}_{\chi^2_d}(1 - \alpha) = z_{d,\alpha}$. Alors la région critique approchée est :

$$W = \{-2 \ln v(x) > z_{d,\alpha}\}.$$

Le test défini par la région critique ci-dessus est considéré comme le test du rapport des vraisemblances maximales dans ce cas. Il n'est pas de seuil exactement égal à α mais seulement asymptotiquement égal à α. □

Ce résultat est aussi valable pour d'autres modèles que les modèles d'échantillon (par exemple pour des cas où les X_i sont indépendantes mais pas de même loi), mais malheureusement pas dans tous les cas.

Chapitre 9

Annexe : tables de lois de probabilité

9.1 Caractéristiques des lois usuelles

9.1.1 Variables aléatoires réelles discrètes

Dans le tableau ci-dessous, on suppose $n \in \mathbb{N}^*$, $p \in]0,1[$ et $\lambda \in \mathbb{R}_+^*$.

Loi et Symbole $X \rightsquigarrow$	Probabilités	$\mathbb{E}[X]$	$\mathbb{V}\text{ar}[X]$	Fonction caractéristique
Bernoulli $\mathcal{B}(p)$	$\mathbb{P}(X=0)=1-p$ $\mathbb{P}(X=1)=p$	p	$p(1-p)$	$1-p+pe^{it}$
Binomiale $\mathcal{B}(n,p)$	$\mathbb{P}(X=k)=\binom{n}{k}p^k(1-p)^{n-k}\mathbb{1}_{\{0,\dots,n\}}(k)$	np	$np(1-p)$	$(1-p+pe^{it})^n$
Binomiale négative $\mathcal{BN}(n,p)$	$\mathbb{P}(X=k)=\binom{k-1}{n-1}p^n(1-p)^{k-n}\mathbb{1}_{\{n,\dots\}}(k)$	$\frac{n}{p}$	$\frac{n(1-p)}{p^2}$	$\left(\frac{pe^{it}}{1-(1-p)e^{it}}\right)^n$
Poisson $\mathcal{P}(\lambda)$	$\mathbb{P}(X=k)=e^{-\lambda}\frac{\lambda^k}{k!}\mathbb{1}_{\mathbb{N}}(k)$	λ	λ	$e^{\lambda(e^{it}-1)}$
Géométrique $\mathcal{G}(p)$	$\mathbb{P}(X=k)=p(1-p)^{k-1}\mathbb{1}_{\mathbb{N}^*}(k)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^{it}}{1-(1-p)e^{it}}$
Hypergéométrique $\mathcal{H}(N,m,n)$ $(m,n)\in\{1,\dots,N\}^2$	$\mathbb{P}(X=k)=\frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}\mathbb{1}_{\{0,\dots,\min(m,n)\}}(k)$	$\frac{nm}{N}$	$\frac{nm(N-n)(N-m)}{N^2(N-1)}$	

9.1.2 Variables aléatoires réelles continues

La fonction Gamma est définie pour $a > 0$ par $\Gamma(a) = \int_0^{+\infty} e^{-x}x^{a-1}dx$.

$$\text{On a : } \forall n \in \mathbb{N}^*, \quad \Gamma(n) = (n-1)!\,, \quad \Gamma(1) = 1\,, \quad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}\,,$$

$$\forall a \in]1, +\infty[\,,\ \Gamma(a) = (a-1)\Gamma(a-1)\,.$$

Dans le tableau ci dessous, $(a, b) \subset \mathbb{R}$, $m \in \mathbb{R}$, $\sigma \in \mathbb{R}_+^*$, $\lambda \in \mathbb{R}_+^*$, $\alpha \in \mathbb{R}_+^*$, $n \in \mathbb{N}^*$

Loi et Symbole $X \rightsquigarrow$	Densité	$\mathbb{E}[X]$	$\mathbb{V}\mathrm{ar}[X]$	Fonction caractéristique
Loi Uniforme $\mathcal{U}(a, b)$	$f_X(x) = \frac{1}{b-a}\mathbb{1}_{[a,b]}(x)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{itb} - e^{ita}}{it(b-a)}$
Loi Normale $\mathcal{N}(m, \sigma^2)$	$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2\sigma^2}}\mathbb{1}_{\mathbb{R}}(x)$	m	σ^2	$e^{itm - \frac{\sigma^2 t^2}{2}}$
Loi Exponentielle $\exp(\lambda) = G(1, \lambda)$	$f_X(x) = \lambda e^{-\lambda x}\mathbb{1}_{\mathbb{R}_+}(x)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\left(1 - \frac{it}{\lambda}\right)^{-1}$
Loi Gamma $G(\alpha, \lambda)$	$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)}e^{-\lambda x}x^{\alpha-1}\mathbb{1}_{\mathbb{R}_+^*}(x)$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$	$\left(1 - \frac{it}{\lambda}\right)^{-\alpha}$
Loi du Chi-deux $\chi_n^2 = G\left(\frac{n}{2}, \frac{1}{2}\right)$	$f_X(x) = \frac{2^{-\frac{n}{2}}}{\Gamma(\frac{n}{2})}e^{-\frac{x}{2}}x^{\frac{n}{2}-1}\mathbb{1}_{\mathbb{R}_+}(x)$	n	$2n$	$(1-2it)^{-\frac{n}{2}}$
Première loi de Laplace	$f_X(x) = \frac{1}{2}e^{-\lvert x\rvert}\mathbb{1}_{\mathbb{R}}(x)$	0	2	$\frac{1}{1+t^2}$

La fonction Beta est définie pour $a > 0$ et $b > 0$ par

$$\beta(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx$$

Dans le tableau suivant, on suppose $a \in \mathbb{R}_+^*,\ b \in \mathbb{R}_+^*, \eta \in \mathbb{R}_+^*\ ,\ \beta \in \mathbb{R}_+^*$.

Loi et Symbole $X \rightsquigarrow$	Densité	$\mathbb{E}[X]$	$\mathbb{V}\mathrm{ar}[X]$
Loi Beta de 1$^{\text{ère}}$ espèce $\beta_1(a,b)$	$f_X(x) = \frac{1}{\beta(a,b)} x^{a-1}(1-x)^{b-1}\mathbb{1}_{[0,1]}(x)$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$
Loi Beta de 2$^{\text{ème}}$ espèce $\beta_2(a,b)$	$f_X(x) = \frac{1}{\beta(a,b)} \frac{x^{a-1}}{(1+x)^{a+b}} \mathbb{1}_{\mathbb{R}_+^*}(x)$	$\frac{a}{b-1}$ si $b > 1$	$\frac{a(a+b-1)}{(b-1)^2(b-2)}$ si $b > 2$
Loi de Weibull $\mathcal{W}(\eta,\beta)$	$f_X(x) = \frac{\beta}{\eta^\beta} x^{\beta-1} e^{-\left(\frac{x}{\eta}\right)^\beta} \mathbb{1}_{\mathbb{R}_+^*}(x)$	$\eta\Gamma\left(1+\frac{1}{\beta}\right)$	$\eta^2\left[\Gamma\left(1+\frac{2}{\beta}\right) - \Gamma\left(1+\frac{1}{\beta}\right)^2\right]$

9.1.3 Vecteurs aléatoires dans $\mathbb{N}^d$ et dans $\mathbb{R}^d$

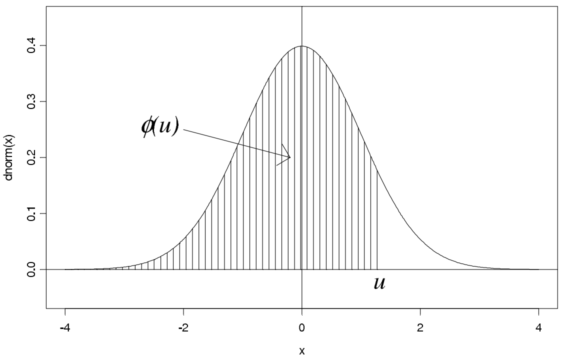
Dans le tableau suivant, on a :
$n \in \mathbb{N}^*, p = (p_1, p_2, \ldots, p_d) \in]0,1[^d$, $\sum_{i=1}^d p_i = 1$, $k = (k_1, k_2, \ldots, k_d) \in \mathbb{N}^d$, $\sum_{i=1}^d k_i = n$.
$m \in \mathbb{R}^d$ et $\Sigma \in M_{d,d}$.

Loi et Symbole $X \rightsquigarrow$	Probabilités ou Densité	$\mathbb{E}[X]$	Matrice de covariance	Fonction Caractéristique
Loi Multinomiale $\mathcal{M}_d(n,p)$	$\mathbb{P}(X=k) = \frac{n!}{k_1!\ldots k_d!} p_1^{k_1} p_2^{k_2} \ldots p_d^{k_d} \mathbb{1}_{\mathbb{N}^d}(k)$	np	$c_{i,i} = np_i(1-p_i)$ $c_{i,j} = -np_ip_j\ ,\ i \neq j$	$\left[\sum_{i=1}^d p_i z_i\right]^n$
Loi normale $\mathcal{N}_d(m,\Sigma)$	$f_X(x) = \frac{1}{\sqrt{det\Sigma}\,(\sqrt{2\pi})^d} e^{-\frac{1}{2}{}^t(x-m)\Sigma^{-1}(x-m)}$	m	Σ	$e^{i^t mt - \frac{1}{2}{}^t t\Sigma t}$

9.2 Tables de lois

9.2.1 Table 1 de la loi normale centrée réduite

U étant une variable aléatoire de loi $\mathcal{N}(0,1)$, la table donne la valeur de $\phi(u) = \mathbb{P}(U \leq u)$. En R, la commande correspondante est `pnorm(u)`.



u	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

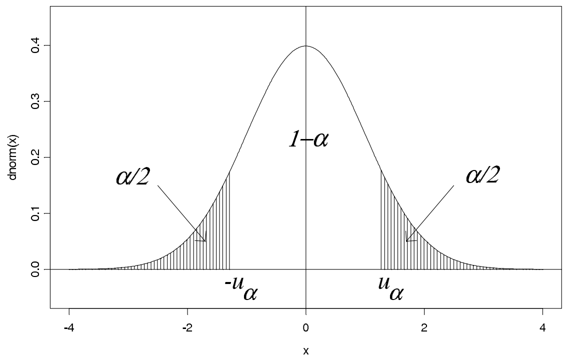
Lecture de la table : $\phi(1.25) = \phi(1.2 + 0.05) = 0.8944$.

Grandes valeurs de u

u	3.0	3.5	4.0	4.5
$\phi(u)$	0.9987	0.99977	0.999968	0.999997

9.2.2 Table 2 de la loi normale centrée réduite

U étant une variable aléatoire de loi $\mathcal{N}(0,1)$ et α un réel de $[0,1]$, la table donne la valeur de $u_\alpha = \phi^{-1}\left(1-\frac{\alpha}{2}\right)$ telle que $\mathbb{P}(|U| > u_\alpha) = \alpha$. En R, la commande correspondante est `qnorm(1-alpha/2)`.



α	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	$+\infty$	2.5758	2.3263	2.1701	2.0537	1.96	1.8808	1.8119	1.7507	1.6954
0.1	1.6449	1.5982	1.5548	1.5141	1.4758	1.4395	1.4051	1.3722	1.3408	1.3106
0.2	1.2816	1.2536	1.2265	1.2004	1.1750	1.1503	1.1264	1.1031	1.0803	1.0581
0.3	1.0364	1.0152	0.9945	0.9741	0.9542	0.9346	0.9154	0.8965	0.8779	0.8596
0.4	0.8416	0.8239	0.8064	0.7892	0.7722	0.7554	0.7388	0.7225	0.7063	0.6903
0.5	0.6745	0.6588	0.6433	0.6280	0.6128	0.5978	0.5828	0.5681	0.5534	0.5388
0.6	0.5244	0.5101	0.4959	0.4817	0.4677	0.4538	0.4399	0.4261	0.4125	0.3989
0.7	0.3853	0.3719	0.3585	0.3451	0.3319	0.3186	0.3055	0.2924	0.2793	0.2663
0.8	0.2533	0.2404	0.2275	0.2147	0.2019	0.1891	0.1764	0.1637	0.1510	0.1383
0.9	0.1257	0.1130	0.1004	0.0878	0.0753	0.0627	0.0502	0.0376	0.0251	0.0125

Lecture de la table : $u_{0.25} = u_{0.2+0.05} = 1.1503$.

Petites valeurs de α

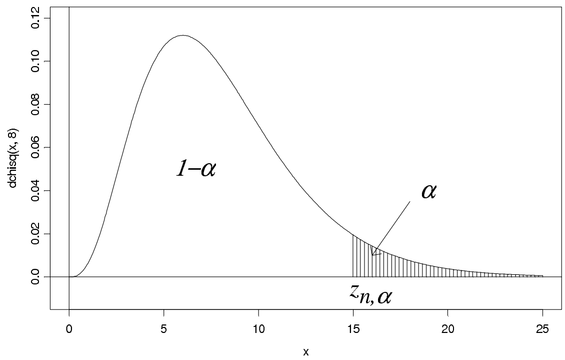
α	0.002	0.001	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-9}
u_α	3.0902	3.2905	3.8906	4.4171	4.8916	5.3267	5.7307	6.1094

$$\text{Pour } p < \frac{1}{2}, \phi^{-1}(p) = -u_{2p}\,.$$

$$\text{Pour } p \geq \frac{1}{2}, \phi^{-1}(p) = u_{2(1-p)}\,.$$

9.2.3 Table de la loi du χ^2

X étant une variable aléatoire de loi du χ^2 à n degrés de libertés et α un réel de $[0,1]$, la table donne la valeur de $z_{n,\alpha} = F^{-1}_{\chi^2_n}(1-\alpha)$ telle que $\mathbb{P}(X > z_{n,\alpha}) = \alpha$. En `R`, la commande correspondante est `qchisq(1-alpha, n)`.



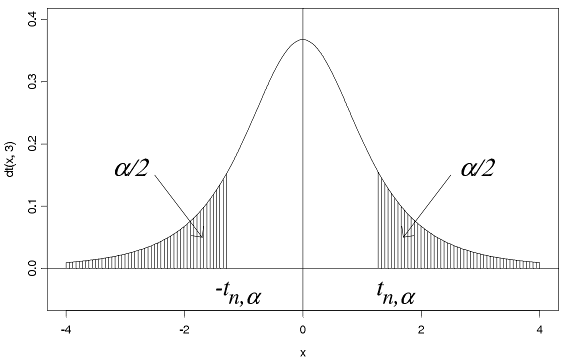
n \ α	0.995	0.990	0.975	0.95	0.9	0.8	0.7	0.5	0.3	0.2	0.1	0.05	0.025	0.01	0.005	0.001
1	0.00004	0.0002	0.001	0.004	0.02	0.06	0.15	0.45	1.07	1.64	2.71	3.84	5.02	6.63	7.88	10.83
2	0.01	0.02	0.05	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.61	5.99	7.38	9.21	10.6	13.82
3	0.07	0.11	0.22	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.81	9.35	11.34	12.84	16.27
4	0.21	0.30	0.48	0.71	1.06	1.65	2.19	3.36	4.88	5.99	7.78	9.49	11.14	13.28	14.86	18.47
5	0.41	0.55	0.83	1.15	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	12.83	15.09	16.75	20.52
6	0.68	0.87	1.24	1.64	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	14.45	16.81	18.55	22.46
7	0.99	1.24	1.69	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	16.01	18.48	20.28	24.32
8	1.34	1.65	2.18	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	17.53	20.09	21.95	26.12
9	1.73	2.09	2.70	3.33	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	19.02	21.67	23.59	27.88
10	2.16	2.56	3.25	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	20.48	23.21	25.19	29.59
11	2.60	3.05	3.82	4.57	5.58	6.99	8.15	10.34	12.90	14.63	17.28	19.68	21.92	24.72	26.76	31.26
12	3.07	3.57	4.40	5.23	6.30	7.81	9.03	11.34	14.01	15.81	18.55	21.03	23.34	26.22	28.30	32.91
13	3.57	4.11	5.01	5.89	7.04	8.63	9.93	12.34	15.12	16.98	19.81	22.36	24.74	27.69	29.82	34.53
14	4.07	4.66	5.63	6.57	7.79	9.47	10.82	13.34	16.22	18.15	21.06	23.68	26.12	29.14	31.32	36.12
15	4.60	5.23	6.26	7.26	8.55	10.31	11.72	14.34	17.32	19.31	22.31	25.00	27.49	30.58	32.80	37.70
16	5.14	5.81	6.91	7.96	9.31	11.15	12.62	15.34	18.42	20.47	23.54	26.30	28.85	32.00	34.27	39.25
17	5.70	6.41	7.56	8.67	10.09	12.00	13.53	16.34	19.51	21.61	24.77	27.59	30.19	33.41	35.72	40.79
18	6.26	7.01	8.23	9.39	10.86	12.86	14.44	17.34	20.60	22.76	25.99	28.87	31.53	34.81	37.16	42.31
19	6.84	7.63	8.91	10.12	11.65	13.72	15.35	18.34	21.69	23.90	27.20	30.14	32.85	36.19	38.58	43.82
20	7.43	8.26	9.59	10.85	12.44	14.58	16.27	19.34	22.77	25.04	28.41	31.41	34.17	37.57	40.00	45.31
21	8.03	8.90	10.28	11.59	13.24	15.44	17.18	20.34	23.86	26.17	29.62	32.67	35.48	38.93	41.40	46.80
22	8.64	9.54	10.98	12.34	14.04	16.31	18.10	21.34	24.94	27.30	30.81	33.92	36.78	40.29	42.80	48.27
23	9.26	10.20	11.69	13.09	14.85	17.19	19.02	22.34	26.02	28.43	32.01	35.17	38.08	41.64	44.18	49.73
24	9.89	10.86	12.40	13.85	15.66	18.06	19.94	23.34	27.10	29.55	33.20	36.42	39.36	42.98	45.56	51.18
25	10.52	11.52	13.12	14.61	16.47	18.94	20.87	24.34	28.17	30.68	34.38	37.65	40.65	44.31	46.93	52.62
26	11.16	12.20	13.84	15.38	17.29	19.82	21.79	25.34	29.25	31.79	35.56	38.89	41.92	45.64	48.29	54.05
27	11.81	12.88	14.57	16.15	18.11	20.70	22.72	26.34	30.32	32.91	36.74	40.11	43.19	46.96	49.64	55.48
28	12.46	13.56	15.31	16.93	18.94	21.59	23.65	27.34	31.39	34.03	37.92	41.34	44.46	48.28	50.99	56.89
29	13.12	14.26	16.05	17.71	19.77	22.48	24.58	28.34	32.46	35.14	39.09	42.56	45.72	49.59	52.34	58.30
30	13.79	14.95	16.79	18.49	20.60	23.36	25.51	29.34	33.53	36.25	40.26	43.77	46.98	50.89	53.67	59.70

$$\text{Pour } n > 30, \text{ on admet que } z_{n,\alpha} \approx \frac{1}{2}\left(u_{2\alpha} + \sqrt{2n-1}\right)^2 \quad \text{si } \alpha < \frac{1}{2}$$

$$\text{et } z_{n,\alpha} \approx \frac{1}{2}\left(\sqrt{2n-1} - u_{2(1-\alpha)}\right)^2 \quad \text{si } \alpha \geq \frac{1}{2}\,.$$

9.2.4 Table de la loi de Student

X étant une variable aléatoire de loi $St(n)$ et α un réel de $[0, 1]$, la table donne la valeur de $t_{n,\alpha} = F^{-1}_{St(n)}\left(1 - \frac{\alpha}{2}\right)$ telle que $\mathbb{P}(|X| > t_{n,\alpha}) = \alpha$. En R, la commande correspondante est `qt(1-alpha/2,n)`. Pour $n = +\infty$, $t_{+\infty,\alpha} = u_\alpha$.



n \ α	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.001
1	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.128	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.768
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.127	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.127	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.126	0.255	0.388	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
80	0.126	0.254	0.387	0.526	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.416
120	0.126	0.254	0.386	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
$+\infty$	0.126	0.253	0.385	0.524	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

Bibliographie

[1] P. Barbe and M. Ledoux, *Probabilité (l3m1)*, Enseignement SUP-Maths, EDP Sciences, 2012.

[2] Patrick Billingsley, *Probability and measure*, 3. ed ed., A Wiley-Interscience publication, Wiley, New York [u.a.], 1995.

[3] Benoît Cadre and Céline Vial, *Statistique mathématique, cours et exercices corrigés*, Ellipses, 2012.

[4] J. Jacod and P. Protter, *L'essentiel en théorie des probabilités*, Collection Enseignement des mathématiques, Cassini, 2003.

[5] Jacques Neveu, *Bases mathématiques du calcul des probabilités*, (No Title) (1964).

[6] Victor Panaretos, *Statistique pour mathématiciens : un premier cours rigoureux*, EPFL Press, 2024.

[7] Gilbert Saporta, *Probabilités, analyse des données et statistique*, Technip, 2011.

[8] C. Wagschal, *Dérivation, intégration*, Collection Méthodes, Hermann, 1999.