Statistiques mathématiques

Equipe pédagogique: A. Barakat, T. Bonald, A. Sabourin, U. Simsekli, G. Staerman

mise à jour: septembre 2019

Table des matières

Ce cours de statistique s'appuie principalement sur les ouvrages de Bickel and Doksum [2015], Lehmann and Casella [1998], Lehmann [1959] et Shao [2008].

Chapitre 1

Analyse statistique des données

1.1 Objectifs de l'analyse statistique, exemples

La plupart des études et des expériences, commerciales, industrielles, ou scientifiques, produisent des données. Au cours de la dernière décennie, le volume total des données stockées a considérablement augmenté, ainsi que les moyens informatiques permettant leur traitement. Une prise de conscience s'opère sur la valeur potentielle de ces grandes masses de données, aussi bien pour le secteur privé que pour le secteur public (par exemple, dans les domaines de la santé publique ou de la gestion des risques industriels, sociétaux ou environnementaux).

L'objet des statistiques est d'extraire de ces données « de la valeur », autrement dit des informations utiles. Le point de vue particulier des statistiques est de considérer ces données comme la réalisation d'une expérience aléatoire. La modélisation mathématique de celle-ci permet de conduire une analyse et un traitement adapté des données (le plus souvent automatique) afin de répondre à des objectifs concrets comme l'apprentissage, le contrôle de qualité, etc. La plupart de ces objectifs particuliers ont un point commun : il s'agit de fournir des outils d'aide à la décision en milieu incertain, en extrayant l'information partielle contenue dans les données à disposition de l'analyste. Dans la suite de ce cours, on utilisera indifféremment les termes *inférence*, *apprentissage*, *analyse statistique* pour faire référence à un processus automatisé d'extraction d'information à partir des données. Avant de formaliser cette approche, donnons quelques exemples.

Exemple 1.1 (Nombre d'objets défectueux)**:**
Considérons une grande population de N éléments, par exemple des objets manufacturés ou des clients d'une entreprise, ou des patients exposés à une maladie. Un nombre inconnu de ces objets, $N\theta$ est défectueux (*resp.* est sur le point de résilier son contrat, c'est-à-dire de « churner », ou est malade). Il est trop coûteux d'examiner individuellement chacun de ces objets. On s'intéresse à la proportion de défauts θ. Pour obtenir une information sur θ, on tire sans remise un échantillon de n éléments parmi N et l'on observe le nombre X d'éléments défectueux (*resp.* de churners, ou de malades) dans cet échantillon. La description mathématique de cet exemple est simple.

Le nombre X d'objets défectueux parmi les n objets choisis au hasard est appelée "observation". L'observation prend donc ici des valeurs entières, positives. Pour n, N et θ fixés, on calcule facilement la loi P_θ :

1. Tout d'abord, X ne "peut pas" valoir plus que n, ni que $N\theta$ (la quantité totale d'objets défectueux). C'est à dire, avec probabilité 1, $X \leq \min(n, N\theta)$.

2. D'autre part, X est positive, et le nombre d'objets non défectueux restants après le tirage, $N(1-\theta)-(n-X)$ est positif. Autrement dit, avec probabilité 1 , $X \geq \max(0, n-N(1-\theta))$.
3. Enfin, pour k un entier entre les deux bornes ci-dessus, la probabilité de choisir k est obtenue par dénombrement : le nombre de choix de k défectueux parmi $N\theta$, multiplié par le nombre de choix de $(n-k)$ non-défectueux parmi les $N - N\theta$ éléments non défectueux, divisé par le nombre total de choix possibles de n éléments parmi N.

On a montré :

$$\mathrm{P}_\theta(\{k\}) = \mathbb{P}(X=k) = \begin{cases} \frac{\binom{N\theta}{k}\binom{N-N\theta}{n-k}}{\binom{N}{n}}, & \text{si } k \in \{\max(n-N(1-\theta),0), \ldots, \min(N\theta, n)\}, \\ 0, & \text{sinon} \end{cases}$$

La loi P_θ définie ci-dessus est appelée *hypergéométrique*, notée $\mathcal{H}yper(N\theta, N, n)$. Cette loi dépend de n, N et θ. La notation P_θ rend compte du fait que θ est un paramètre inconnu qui détermine (une fois fixés N et n) la loi de X. Dans cet exemple, la description de l'expérience aléatoire produisant l'observation nous a permis de spécifier la loi de probabilité de l'observation à l'inconnue θ près. Autrement dit, notre connaissance sur cette loi est qu'elle appartient à une famille

$$\Big\{\mathrm{P}_\theta = \mathcal{H}yper(N\theta, N), \theta \in \{0, \frac{1}{N}, \frac{2}{N}, \ldots, 1\}\Big\}.$$

L'expérience nous fournira une information permettant par exemple d'estimer la valeur de θ. Par exemple, on peut montrer que l'espérance de X vaut $n\theta$. Un estimateur "raisonnable" de θ (au sens où l'estimation est "en moyenne juste", c'est-à-dire "non-biaisée"), est $\widehat{\theta} = X/n$. L'estimateur est bien une fonction des données.

Exemple 1.2 (Modèle à deux échantillons, test A/B)**:**
Soient $X = (X_1, \ldots, X_m)$ et $Y = (Y_1, \ldots, Y_n)$ les réponses respectivement de m sujets ayant une pathologie particulière à un traitement A et de n sujets souffrant de la même pathologie à un traitement B. Par convention, A est un traitement standard ou un placebo et X est la population de dite de *contrôle*. Un placebo est une substance dont on est sûr qu'il n'a pas d'effet sur la pathologie considéré, et est utilisé pour corriger l'effet "placebo". Y représente les réponses des patients à un nouveau traitement, dont on évalue l'effet par rapport au placebo. On appelle Y l'observation de la population test. Dans le cadre du marketing, A est un produit ou une page web standard, alors que B est une nouvelle version, dont on cherche à déterminer l'effet sur les consommateurs en soumettant la population de contrôle X à une version standard alors qu'on propose B à la population test Y.

Les hypothèses naturelles sont

(i) Les v.a. $X_1, \ldots, X_m$ sont *i.i.d.* (indépendantes et identiquement distribuées) de loi F et $Y_1, \ldots, Y_n$ sont i.i.d. de loi G, indépendantes de X. La loi jointe de toutes les observations est donc spécifiée par la donnée de la paire (F, G),

(ii) Une hypothèse souvent faite est celle de la *constance de l'effet du traitement*. Supposons que le traitement A soit administré à un patient, et que la réponse x soit obtenue. L'hypothèse de la constance de l'effet de traitement consiste à dire que si le traitement B avait été administré à ce même patient, alors la réponse $y = x + \Delta$ aurait été obtenue, où Δ *ne dépend pas de* x. En terme probabiliste, ceci signifie que si F est la loi de la population de contrôle, alors la loi de la distribution de test est $G(\,\cdot\,) = F(. - \Delta)$. Nous appellerons de tels modèles des modèles de *translation*.

(iii) Une autre hypothèse simplificatrice peut être faite. On peut supposer par exemple que la loi F de la population de contrôle est une loi normale de moyenne μ et de variance σ^2, $F = \mathcal{N}(\mu, \sigma^2)$. Sous l'hypothèse précédente, $G = \mathcal{N}(\mu + \Delta, \sigma^2)$. Ce modèle, très classique, est le modèle à deux échantillons gaussiens, de même variance.

L'analyse statistique aura alors pour but, par exemple de déterminer (toujours au vu des données) si Δ est significativement différent de 0 ou non (cadre des tests statistiques, que nous verrons dans un chapitre ultérieur), ou encore d'estimer la valeur de Δ (cadre de l'estimation ponctuelle), ou de déterminer si Δ est plus grand qu'un certain seuil réglementaire δ_0 fixé (à nouveau, cadre d'un test statistique).

L'exemple 1.2 montre que plusieurs modèles sont envisageables pour une même expérience aléatoire. D'où la question du *choix du modèle*. Ce qui fait un bon modèle est un mélange d'expérience, de connaissance a priori, de considération sur les lois physiques (ou économiques, biologiques, ...) ayant engendré les données et bien sûr d'hypothèses de travail. Une spécification très précise de la structure du modèle permet en général de réduire la partie inconnue du modèle (les paramètres μ, Δ et σ^2 dans l'exemple 1.2 sous l'hypothèse *(iii)*), ce qui simplifie les procédures d'estimation de grandeurs d'intérêt dépendant de la loi inconnue des observations. Cependant, si le modèle est mal spécifié, nos analyses, bien que correctes sur le plan mathématique, peuvent conduire à des interprétations fausses des estimations produites.

1.2 Formalisation statistique d'un problème

Généralisons les exemples précédents :

1.2.1 Cadre probabiliste, notations

Un rappel succinct des éléments et des notations indispensables de théorie de la mesure et de l'intégration est donné en annexe (chapitre A).

Donnons-nous tout d'abord un univers Ω, un ensemble non vide décrivant l'ensemble des réalisations possibles de l'expérience. Un élément $\omega \in \Omega$ est une *réalisation* (ou *épreuve*) particulière. Par exemple, dans l'exemple 1.1, on peut prendre comme espace Ω l'ensemble $\{0,1\}^n$ ou $\{\mathrm{D},\mathrm{N}\}^n$ (D : objet défectueux ; N : objet non-défectueux) ;

Malheureusement l'ensemble des réalisations Ω n'est pas toujours aussi simple (fini ou dénombrable). Une expérience décrite par un nombre réel quelconque, $\Omega = \mathbb{R}$, une mesure d'une quantité numérique par exemple ne se décrit pas par un ensemble dénombrable de possibilités. On introduit donc la notion d'*événement* : un événement est un sous-ensemble particulier de Ω. L'ensemble des *événements* que l'on notera $\mathcal{F}$, aura la structure d'une tribu, on appellera donc cet ensemble $\mathcal{F}$ la *tribu des événements*.[1] .

Pour la modélisation statistique, nous nous concentrons souvent sur certaines quantités résumant l'issue de l'expérience : dans l'exemple 1.1, on s'intéresse seulement au nombre d'objets défectueux et non pas à l'ordre dans lequel les objets défectueux apparaissaient dans l'échantillon. Pour prendre en compte ce fait, on construit

1. un espace d'*observations* $\mathcal{X}$, a priori distinct de l'espace des épreuves Ω, que nous munissons d'une tribu $\mathcal{B}(\mathcal{X})$, composée de parties de $\mathcal{X}$;

1. La notion de tribu impose des propriétés minimales de stabilité pour $\mathcal{F}$ nécessaires au calcul des probabilités de ces ensembles. Pour la compréhension de ce chapitre, on peut supposer que la *tribu des événements* est tout simplement l'ensemble des parties de Ω.

2. une variable aléatoire X (appelée *observation*) définie sur l'espace des épreuves $(\Omega, \mathcal{F})$ et à valeurs dans l'espace des observations $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, c'est-à-dire une fonction mesurable $X : (\Omega, \mathcal{F}) \to (\mathcal{X}, \mathcal{B}(\mathcal{X}))$.

Dans l'exemple 1.1, l'espace des observations est $\mathcal{X} = \{0, 1, \ldots, n\}$, à savoir le nombre d'objets défectueux dans un échantillon de n objets ; alors que l'ensemble des événements est $\Omega = \{0, 1\}^n$. Comme Ω et $\mathcal{X}$ sont dénombrables, nous munissons ces ensembles des tribus de toutes leurs parties, $\mathcal{F} = \mathcal{P}(\Omega)$ et $\mathcal{B}(\mathcal{X}) = \mathcal{P}(\mathcal{X})$. La variable aléatoire X est alors donné par $X(\omega_1, \ldots, \omega_n) = \sum_{i=1}^n \mathbb{1}\{\omega_i = 0\}$, où $(\omega_1, \ldots, \omega_n) \in \{0, 1\}^n$.

Dans certaines situations, il n'est pas nécessaire de distinguer l'espace des épreuves Ω et l'espace des réalisations $\mathcal{X}$. Dans ce cas, on posera $(\Omega, \mathcal{F}) = (\mathcal{X}, \mathcal{B}(\mathcal{X}))$, et on prendra simplement $X(\omega) = \omega$ pour tout $\omega \in \Omega$.

Remarquons que, jusqu'à présent, on n'a pas introduit de loi de probabilité $\mathbb{P}$ sur $(\Omega, \mathcal{F})$ ni de loi P sur $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ selon laquelle X serait générée. En effet, en statistique, une telle loi sous-jacente est inconnue et l'objectif général de l'analyse statistique est d'extraire une information de l'observation X concernant la loi de probabilité qui l'a générée.

1.2.2 Modèle statistique et paramétrisation

En statistiques il n'est pas question de comprendre exactement *comment* l'observation X a été générée. En revanche il s'agit de comprendre le mieux possible quelle est sa *loi*. Cette connaissance provient d'une part d'une connaissance *a priori* et d'autre part du résultat d'une expérience aléatoire. La connaissance a priori est formalisée par la donnée d'une famille $\mathcal{P}$ de probabilités sur l'espace des observations $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. La famille $\mathcal{P}$ sera appelée le *modèle statistique* pour le problème considéré. Dans l'exemple 1.1, le modèle $\mathcal{P}$ est la famille des lois hypergéométriques de paramètre θ pour un échantillon de taille n d'une population N. On verra plus tard, au chapitre concernant la statistique bayésienne, qu'on peut aller plus loin dans la formalisation de la connaissance a priori.

Il est souvent pratique de définir une *paramétrisation* du modèle, c'est-à-dire d'étiqueter chaque loi $\mathrm{P} \in \mathcal{P}$ par un *paramètre* $\theta \in \Theta$, où Θ est un ensemble quelconque appelé *espace des paramètres*. On écrira alors P_θ pour désigner la loi ainsi étiquetée. On choisira en particulier Θ de sorte que la loi P_θ soit entièrement déterminée par le paramètre θ. Formellement, une paramétrisation de $\mathcal{P}$ est une application $\theta \mapsto \mathrm{P}_\theta$ définie de *l'espace des paramètres* Θ dans l'ensemble $\mathcal{P}$, surjective (chaque loi P doit pouvoir être étiquetée). Dans l'exemple introductif 1.1, si l'on fixe N et n, la loi P de X est entièrement déterminée par θ. On peut donc écrire $\mathrm{P}_\theta = \mathcal{H}yper(N\theta, N)$. L'ensemble des lois possibles des observations est donc $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$ où l'ensemble des paramètres Θ est $\{0, 1/N, \ldots, 1\}$.

Définition 1.2.1 (Modèle statistique, espace des paramètres)**.** *Nous appelons* modèle statistique *une famille de probabilités $\mathcal{P}$ sur* l'espace des observations $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. *Si Θ est un ensemble quelconque tel que*

$$\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\},$$

alors Θ est appelé espace des paramètres *du modèle.*

Remarque 1.2.2. *(Existence) Remarquons qu'il est toujours possible de paramétrer un ensemble par lui-même,* via *l'application identité. On pourra donc toujours définir un espace des paramètres Θ, quitte à prendre $\Theta = \mathcal{P}$, ce qui ne présente pas beaucoup d'intérêt mais nous permettra d'écrire systématiquement les modèles considérés $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$ sans avoir besoin de se poser la question de l'existence d'une telle paramétrisation.*

Le résultat d'une expérience aléatoire est alors interprété comme étant la réalisation d'une variable aléatoire X à valeurs dans $\mathcal{X}$ et de loi P_θ appartenant au modèle statistique $\mathcal{P}$, c'est-à-dire telle que $\theta \in \Theta$. La variable X s'appelle l'*observation* (ou encore la donnée, les données, ...). Dans la suite de ce cours, la notation « $X \sim \mathrm{P}_\theta$ » signifie

« La variable aléatoire X est distribuée selon la loi P_θ ».

Le travail du statisticien peut se décrire ainsi :

- **La seule connaissance mise à la disposition du statisticien est un modèle $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ et une réalisation de l'observation $X \sim \mathrm{P}_\theta$, où $\theta \in \Theta$ est inconnu.**
- **L'objectif est d'approcher une certaine quantité d'intérêt $g(\theta)$ (dépendant uniquement de θ) en utilisant une procédure fondée uniquement sur l'observation X (une fonction ne dépendant que de X).**

Autrement dit, le statisticien est amené à proposer des méthodes construites à partir de *fonctions des données*. Ceci mène à la notion de *statistique*, qui a un sens précis donné dans la définition 1.2.3 ci-dessous. Rappelons que si φ est une fonction mesurable définie sur $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ à valeurs dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, alors $\varphi(X)$ est encore une variable aléatoire (en effet, la fonction $\varphi \circ X$ est mesurable de $(\Omega, \mathcal{F})$ dans $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$).

Définition 1.2.3. *Une* statistique *est une variable aléatoire s'écrivant comme une fonction mesurable des observations, de type $\varphi(X)$ où $\varphi : (\mathcal{X}, \mathcal{B}(\mathcal{X})) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ est une fonction mesurable.*

Ainsi, une statistique est une fonction mesurable quelconque des observations.

Quand il sera nécessaire d'utiliser la v.a. X, définie sur $(\Omega, \mathcal{F})$ et de loi P_θ, dans les calculs, on utilisera la notation $\mathbb{P}_\theta$ et $\mathbb{E}_\theta$ pour la probabilité définie sur $\mathcal{F}$ et l'espérance associée, par exemple,

$$\mathbb{P}_\theta(X \in A) = \mathrm{P}_\theta(A) \quad \text{et} \quad \mathbb{E}_\theta[\varphi(X)] = \int_{\mathcal{X}} \varphi(x) \mathrm{P}_\theta(\mathrm{d}x) \,, \tag{1.1}$$

pour $A \in \mathcal{B}(\mathcal{X})$ et φ est une fonction mesurable telle que l'intégrale est correctement définie.

1.3 Modèles paramétriques, non-paramétriques ; identifiabilité.

Considérons un modèle statistique de la forme $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$. Lorsque Θ peut être choisi comme sous-ensemble d'un espace euclidien (de dimension finie), le modèle sera dit *paramétrique*. Sinon, on dira que le modèle est *non–paramétrique*. Enfin, si Θ est inclus dans un espace de la forme $\Theta_1 \times \Theta_2$ où Θ_1 est inclus dans un espace euclidien, alors on dira que le modèle est *semi–paramétrique*.

Exemple 1.3 (modèle fini)**:**
Dans l'exemple 1.1, la loi des observations est entièrement déterminée par la proportion θ d'objets défectueux. On peut donc paramétrer le modèle par $\Theta = \{0, 1/N, \ldots, 1\} \subset \mathbb{R}$ et noter $\mathrm{P}_\theta = \mathcal{H}yper(N\theta, N, n)$. Le modèle est paramétrique, et même fini puisque le nombre de valeurs possibles pour θ est fini.

Les exemples 1.4, 1.5 et 1.6 ci-dessous introduisent deux modèles de cardinal infini (nombre infini de θ possibles), l'un paramétrique, l'autre non-paramétrique.

Supposons que nous cherchions à déterminer comment une grandeur physique ou économique, par exemple, la taille ou les revenus, est distribuée dans une grande population. Un recensement exhaustif est trop coûteux, et ces quantités doivent donc être mesurées par sondage en choisissant au hasard un échantillon de taille n de cette population. Il s'agit d'un problème similaire au précédent (exemple 1.3), à la différence que nous mesurons cette fois un attribut numérique (taille, revenu) plutôt qu'un nombre entier. Une épreuve est un vecteur $\omega = (\omega_1, \ldots, \omega_n)$ de valeurs réelles, et nous poserons donc ici $\Omega = \mathbb{R}^n$ et $\mathcal{F} = \mathcal{B}(\mathbb{R}^n)$. Il n'y a pas lieu de distinguer ici l'espace des observations et des épreuves et nous poserons donc $\mathcal{X} = \Omega$, $\mathcal{B}(\mathcal{X}) = \mathcal{F}$ et $X = (X_1, \ldots, X_n)$ avec $X_i(\omega) = \omega_i$ pour tout $i \in \{1, \ldots, n\}$ et $\omega \in \Omega$. Si nous supposons que les attributs numériques sont indépendants, de même loi F, la loi de l'observation X est égale au produit tensoriel des lois F, *i.e.* pour toute suite $A_1, \ldots, A_n$ de boréliens,

$$\mathrm{P}_\theta(A_1 \times \cdots \times A_n) = \mathbb{P}_\theta(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} F(A_i).$$

Différentes approches peuvent être considérées.

Exemple 1.4 (Observation numérique, modèle paramétrique)**:**
Nous pouvons par exemple supposer que F est une loi normale, de moyenne et de variance inconnue, *i.e.* $F = \mathcal{N}(\mu, \sigma^2)$ avec $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}^+$, où $\mathcal{N}(\mu, \sigma^2)$ est la loi d'une v.a. gaussienne de moyenne μ et de variance σ^2. Posons alors $\theta = (\mu, \sigma^2)$, et $\Theta = \mathbb{R} \times]0, \infty[$. Rappelons que la Gaussienne de paramètre θ a pour densité :

$$\phi(x; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Ainsi, pour tout $\theta \in \Theta$, définissons la loi P_θ comme la loi Gaussienne produit sur $\mathbb{R}^n$ (*i.e.* la loi d'un vecteur gaussien de composantes indépendantes) de densité marginale $\phi(\,\cdot\,, \theta)$. Ainsi, P_θ est définie par :

$$\mathrm{P}_\theta(A) = \int_A \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) dx_1 \ldots dx_n, \quad \forall A \in \mathcal{B}(\mathbb{R}^n).$$

Ici, comme dans le cas précédent, la loi des observations est, entièrement déterminée par le paramètre $\theta \in \Theta \subset \mathbb{R}^2$. C'est donc un *modèle paramétrique*. Bien sûr ces paramètres sont inconnus, et un des objets de l'inférence sera de déterminer (ou plutôt d'approcher) θ en utilisant l'information contenue dans les données.

Exemple 1.5 (Observation numérique, modèle non-paramétrique)**:**
Une autre approche, reposant sur moins d'information a priori, consiste à supposer que F est une loi admettant une densité f régulière (par exemple deux fois différentiable sur $\mathbb{R}$). Une telle approche est *non–paramétrique*. Bien que non–paramétrique, notons toutefois que nous avons déjà formulé des hypothèses sur le mécanisme de génération des données, en particulier que les observations sont indépendantes et identiquement distribuées et que la loi F admet une densité régulière.

Exemple 1.6 (Obervation numérique, modèle dit « semi-paramétrique »)**:**
Une approche intermédiaire consiste par exemple à supposer que

(i) la loi F admet une densité $f(\cdot - \mu)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$, $F(A) = \int_A f(x-\mu)dx$,
(ii) la densité f est symétrique sur $\mathbb{R}$: $f(x) = f(-x)$.

Par rapport à la première situation, l'hypothèse faite est a priori plus faible, car nous ne spécifions pas la densité f (nous imposons simplement qu'elle soit symétrique) et nous privilégions un paramètre *d'intérêt* μ. C'est un modèle *semi-paramétrique*. L'ensemble des paramètres est

$$\Theta = \{(\mu, f) \ : \mu \in \mathbb{R}, f \text{ densité symétrique}\} \tag{1.2}$$

Ainsi, le modèle statistique est $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$ où la loi P_θ est définie par :

$$\mathrm{P}_\theta(A) = \int_A \prod_{i=1}^n f(x_i - \mu)\, dx_1 \dots dx_n, \quad \forall A \in \mathcal{B}(\mathbb{R}^n)\ ,$$

Remarquons qu'il existe en général de multiples manières de définir une paramétrisation. N'importe quelle transformation bijective sur Θ permet en particulier de définir une nouvelle paramétrisation. Par exemple, nous pourrions choisir de paramétrer la loi gaussienne par $(\mu, \mu^2 + \sigma^2)$ plutôt que par (μ, σ^2). La paramétrisation que nous choisissons est en général naturellement dictée par le phénomène que nous modélisons, bien que la paramétrisation qui semble la plus naturelle ne soit pas toujours nécessairement celle qui se prête le mieux à l'analyse mathématique. Un problème important pour le choix d'une paramétrisation est celui de l'*identifiabilité*.

Définition 1.3.1 (Identifiabilité)**.** *Un modèle statistique $\mathcal{P}$ décrit par un paramètre $\theta \in \Theta$, $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$, est dit* identifiable *si, pour tout θ_1 et θ_2 de Θ, l'égalité $\mathrm{P}_{\theta_1} = \mathrm{P}_{\theta_2}$ implique $\theta_1 = \theta_2$.*
Plus généralement, une fonction $g(\theta)$ du paramètre θ est dite identifiable si l'égalité $\mathrm{P}_{\theta_1} = \mathrm{P}_{\theta_2}$ implique $g(\theta_1) = g(\theta_2)$.

Autrement dit, le paramètre est identifiable si l'application $\theta \mapsto \mathrm{P}_\theta$ est injective. Dans l'exemple 1.6, supposons que nous remplacions l'ensemble des paramètres (1.2) par l'ensemble plus grand :

$$\Theta = \{(\mu, f) \ : \mu \in \mathbb{R}, f \text{ densité}\}\ ,$$

c'est-à-dire qu'on ne restreint plus f aux densités symétriques. Cette paramétrisation n'est pas identifiable, par exemple, nous pouvons prendre $\mu = 0$ et f égale à la densité de la loi $\mathcal{N}(0,1)$ ou $\mu = 1$ et f égale à la densité de la loi $\mathcal{N}(-1,1)$.

Remarque 1.3.2 (Existence d'un espace de paramètres identifiable)**.** *Pour conclure sur la notion de paramétrisation et d'identifiabilité, notons qu'il existe* toujours *une paramétrisation $\{\mathcal{P}_\theta, \theta \in \Theta\}$ qui soit identifiable : il suffit de prendre $\Theta = \mathcal{P}$ et $\theta = \mathrm{P}$. Ceci ne présente pas d'intérêt pratique pour la modélisation mais permet d'utiliser la notation P_θ sans avoir à se poser la question de l'existence d'une telle paramétrisation ou de son identifiabilité.*

Il est courant en pratique de parler d'un *paramètre* sans supposer que ce paramètre caractérise entièrement la loi. En effet, on peut être intéressé par certaines caractéristiques particulières d'une loi (son espérance par exemple), sans vouloir la connaître tout entière. Ceci se formalise en définissant une application g de l'espace Θ dans un espace $\mathcal{G}$ quelconque (les valeurs possibles prises par la grandeur d'intérêt). Un paramètre $g(\theta)$ est alors une caractéristique de la distribution P_θ. Dans l'exemple 1.6, la quantité μ pourra être appelée paramètre (de localisation), même si elle ne détermine pas entièrement la loi.

Exemple 1.7 (Modèles de régression)**:**
Soit $X_i = (Z_i, Y_i) \in \mathcal{X} = \mathbb{R} \times \mathbb{R}^p$, $i = 1, \dots, n$ un échantillon vérifiant

$$Y_i = \langle \theta, Z_i \rangle + \theta_0 + \xi_i \text{ pour tout } i \in \{1, \dots, n\}, \tag{1.3}$$

où $(\xi_1, \dots, \xi_n)$ et $(Z_1, \dots, Z_n)$ sont deux échantillons i.i.d. indépendants respectivement de loi $\mathrm{P}^{(b)}$ et $\mathrm{P}^{(r)}$, θ est un paramètre dans $\mathbb{R}^p$ et $\theta_0 \in \mathbb{R}$. On suppose que l'on observe $X_1, \dots, X_n \in \mathcal{X}$. Le modèle est entièrement spécifié par la donnée du paramètre $\tilde{\theta} = (\theta, \theta_0)$, de la loi $\mathrm{P}^{(b)}$ et de la loi $\mathrm{P}^{(r)}$. Les variables Z_i, $i = 1, \dots, n$ sont appelés les *régresseurs* (ou valeur explicatives) du modèle. Ici, l'observation Y dépend de manière affine du régresseur Z_i, à un bruit additif ξ près. On parle alors de modèle de *régression linéaire*.

Plus généralement, on aurait pu simplement supposer que l'observation est fonction des régresseur et du bruit, *i.e.*

$$Y_i = f(Z_i, \xi_i) \text{ pour tout } i \in \{1, \dots, n\}\,, \tag{1.4}$$

où le « paramètre » f est une fonction $\mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$. Si l'on suppose que $f(z, \xi) = g(z) + \xi$, on parlera de *modèle de bruit additif*. Ainsi, un modèle de régression linéaire est un modèle de régression avec bruit additif et où g est affine.

Dans le modèle de régression (1.4), les *paramètres d'intérêt* sont ceux qui décrivent la fonction f. Dans le cas linéaire (1.3), le paramètre d'intérêt est $\tilde{\theta}$. La loi $\mathrm{P}^{(b)}$ est en général inconnue mais pas nécessairement. Les paramètres inconnus qui la déterminent sont appelés *paramètres de nuisance* . La loi $\mathrm{P}^{(r)}$ est en générale inconnue. Elle n'est pas cruciale pour définir des procédures d'estimation puisque ces variables sont observées (contrairement aux ξ_i, $i = 1, \dots, n$). Toutes ces procédures peuvent en effet être décrites en fonction de ces variables en les considérant comme des variables déterministes.

Enfin, dans le cadre des modèles de régression, on appelle *prédicteur* une fonction $h : \mathbb{R}^d \to \mathbb{R}$ qui permet d'estimer Y à partir de Z, par exemple, la fonction définie par $h(z) = \mathbb{E}[f(z, \xi)]$, où $\xi \sim \mathrm{P}^{(b)}$. Si le prédicteur est de la forme $h(z) = \langle T, z \rangle + T_0$ (une fonction affine de z) on parle de *prédicteur linéaire*. Un problème important de l'estimation en régression est de trouver un *prédicteur estimé* à partir d'observations $X_1, \dots, X_n$.

Dans cet exemple on a vu les notions de *paramètre d'intérêt* et *paramètre de nuisance*. Ces notions générales expriment une hiérarchie dans l'importance des paramètres du point de vue pratique.

1.4 Modèles dominés

On parlera de *modèle dominé* $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$ lorsque toutes les lois $\mathrm{P}_\theta \in \mathcal{P}$ admettent une densité par rapport à une *même* mesure de référence μ. Le cas le plus fréquent est celui où le modèle est dominé par la mesure de Lebesgue sur $\mathbb{R}^n$. Alors la famille de loi est définie directement par la donnée d'une famille de densités de probabilité par rapport à une mesure sous-jacente (le plus souvent, la mesure de Lebesgue multi-dimensionnelle).

Définition 1.4.1 (Modèle dominé)**.** *Nous dirons qu'un modèle statistique $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$ est* dominé *s'il existe une mesure positive μ sur $\mathcal{B}(\mathcal{X})$ telle que pour tout $\theta \in \Theta$, $\mathrm{P}_\theta \in \mathcal{P}$ admette une densité de probabilité p_θ par rapport à μ.*

Remarque 1.4.2. *Le fait d'admettre une densité par rapport à une mesure donnée est intimement lié à la notion de* relation de domination *entre des mesures positives, rappelée dans la section A.5. En effet, le théorème de Radon-Nikodym (voir le théorème A.5.2) assure que pour deux mesures positives fixées* P *et* μ*,* P *admet une densité par rapport à* μ *si et seulement si* P *est* absolument continue *par rapport à* μ*(on note* $\mathrm{P} \ll \mu$*), c'est-à-dire si pour tout ensemble mesurable* A*,* $\mu(A) = 0 \Rightarrow \mathrm{P}(A) = 0$*.*

Les cas suivants sont les plus courants :

1. Le modèle $(\mathrm{P}_\theta, \theta \in \Theta)$ est dominé par la mesure de Lebesgue sur $\mathbb{R}^d$,

$$\mathrm{P}_\theta(A) = \int_A p_\theta(x)dx, \quad \text{et} \quad \mathbb{E}_\theta[\varphi(X)] = \int_{\mathbb{R}^d} \varphi(x)p_\theta(x)dx.$$

   pour $A \in \mathcal{B}(\mathbb{R}^d)$ et φ une fonction borélienne positive ou bornée.

2. L'espace $\mathcal{X}$ est fini ou dénombrable, et le modèle $(\mathrm{P}_\theta, \theta \in \Theta)$ est dominé par la mesure de comptage sur $\mathcal{X}$,

$$\mathrm{P}_\theta(A) = \sum_{x \in A} p_\theta(x), \quad \text{et} \quad \mathbb{E}_\theta[\varphi(X)] = \sum_{x \in \mathcal{X}} \varphi(x)p_\theta(x) \quad \text{où} \quad p_\theta(x) = \mathbb{P}_\theta[X = x].$$

Dans l'exemple 1.1, le modèle $\{\mathrm{P}_\theta, \theta \in \Theta\}$ est dominé par la mesure μ de comptage sur $\mathcal{X} = \{0, \ldots, n\}$.

Remarque 1.4.3. *Tout modèle défini sur un espace fini ou dénombrable* $(\mathcal{X}, \mathcal{P}(\mathcal{X}))$ *est dominé par la mesure de comptage sur* $\mathcal{X}$

$$\mu = \sum_{x \in \mathcal{X}} \delta_x \ .$$

Par définition, tout modèle dominé (paramétré par Θ) est entièrement caractérisé par la famille de densités $\{p_\theta, \theta \in \Theta\}$. L'intérêt est donc de pouvoir travailler directement sur une famille de densités au lieu d'une famille de mesures de probabilité. Ceci permet d'utiliser la notion de *vraisemblance* définie ci-dessous. Nous noterons alors $p(.;\theta)$ ou $p_\theta(.)$, suivant le contexte, la densité de la loi P_θ par rapport à une mesure dominante de référence μ.

Définition 1.4.4 (Vraisemblance)**.** *L'application* $\theta \to p(x;\theta)$ *s'appelle la* fonction de vraisemblance *de l'observation* x*.*

La vraisemblance est l'ingrédient de base d'une large famille de procédures d'inférence (appelées justement méthodes de vraisemblance, ou méthodes basées sur le principe de vraisemblance), dont nous verrons quelques exemples plus loin dans ce cours (méthodes bayésiennes, estimateur du maximum de vraisemblance).

Une première interprétation intuitive est la suivante : étant donné une observation donnée x, il est d'autant plus « vraisemblable » que l'observation ait été générée sous la loi P_θ que la valeur de la densité $p(x;\theta)$ est élevée (d'où le terme de « vraisemblance »). Ainsi, la vraisemblance peut être vue comme une « note » attribuée au paramètre θ : plus la note est élevée, plus il est raisonnable de penser que c'est bien P_θ qui est à l'origine de l'observation x. Cette heuristique est à la base des estimateurs obtenus par maximisation de la vraisemblance (voir le chapitre 2).

1.5 Nombre d'observations

Jusqu'à maintenant, notre description ne prend pas en compte une notion importante de la modélisation statistique : le *nombre d'observations* ; elle décrit un modèle à *"n fixé"*. Dans les exemples proposés nous voyons immédiatement qu'il existe explicitement un nombre d'observations dans la modélisation : le nombre d'éléments n dans l'échantillon dans les exemples 1.1 et 1.5, le couple (m, n) dans l'exemple 1.2. L'objectif des statistiques asymptotiques abordées dans un cours ultérieur est de comprendre comment les procédures statistiques évoluent quand ce nombre devient grand. Pour l'instant, contentons-nous de comprendre comment le nombre d'observations peut intervenir dans la description d'un modèle. Pour ce faire nous écrirons momentanément le modèle statistique sous la forme $\mathcal{P}_n$ et un élément de ce modèle sous la forme P_n. Il arrivera souvent que le modèle $\mathcal{P}_n$ dépende uniquement d'un modèle plus simple $\mathcal{P}$ (en général $\mathcal{P} = \mathcal{P}_1$) et de n. Le cadre le plus simple est celui d'un échantillon *i.i.d.* (indépendant et identiquement distribué). On dira alors que $(X_1, \ldots, X_n)$ est un échantillon i.i.d. de loi $\mathrm{P} \in \mathcal{P}$ sur $\mathcal{X}$, ce qui signifie, dans ce cas, que l'observation $X = (X_1, \ldots, X_n) \in \mathcal{X}^n$ est de loi

$$\mathrm{P}_n = \mathrm{P}^{\otimes n} \quad \text{(loi produit)},$$

où P est une probabilité sur $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. On rappelle que, étant donné une loi P sur $\mathbb{R}$, la loi produit $\mathrm{P}^{\otimes n}$ définie sur $\mathbb{R}^n$ est donnée par

$$\mathrm{P}^{\otimes n}(A_1 \times \ldots \times A_n) = \prod_{i=1}^{n} \mathrm{P}(A_i), \qquad A_1, \ldots, A_n \subset \mathbb{R}.$$

Dans ce cas, le modèle à n observations est donné par

$$\mathcal{P}_n = \left\{ \mathrm{P}^{\otimes n} \ : \ \mathrm{P} \in \mathcal{P} \right\} ,$$

et on a bien $\mathcal{P}_1 = \mathcal{P}$. C'est l'hypothèse faite dans l'exemple 1.5 mais aussi, à peu de modifications près, dans l'exemple 1.2, en adaptant la relation ci-dessus sous la forme $\mathrm{P}_{F,G,n} = F^{\otimes m} \otimes G^{\otimes n}$. L'espace $\mathcal{X}$ est souvent de dimension un mais pas toujours. En particulier, s'il est de dimension supérieure, cette liberté laisse place à la modélisation de données dépendantes. Par exemple, un modèle i.i.d. de *vecteurs gaussiens* supposera $\mathcal{X} = \mathbb{R}^d$ et $X_i \sim \mathrm{P}_\theta = \mathcal{N}(\mu, \Sigma)$ avec $\theta = (\mu, \Sigma)$ avec $\mu \in \mathbb{R}^d$ et Σ une matrice $d \times d$ symétrique positive.

1.6 Actions, procédures de décision, fonction de perte et risque

Étant donné un modèle statistique, l'information que nous voulons tirer des observations varie suivant les objectifs de notre analyse. Nous pouvons par exemple chercher à découvrir les valeurs des paramètres importants, par exemple, la proportion des objets défectueux dans l'exemple 1.1 ou de la constante μ dans l'exemple 1.5. On parle alors de problèmes d'*estimation*.

L'estimation n'est pas le seul problème que l'on peut être amené à se poser : dans l'exemple 1.2, une question possible est de déterminer si la distribution F de la population de référence est significativement différente de la distribution G de la population de test, ou, sous l'hypothèse de constance de l'effet de traitement, que $\Delta \neq 0$ (le traitement est efficace). Il s'agit ici d'un problème de *test statistique*, où nous cherchons à déterminer si deux distributions sont différentes, ou, dans un cadre paramétrique, si la valeur d'un paramètre

excède un certain seuil. Le type de réponse que l'on attend d'une procédure d'estimation ou de test, nous dirons, plus généralement, d'une *procédure de décision*, s'appelle une *action*.

Nous appellerons donc l'*espace des actions* $\mathcal{A}$, les valeurs prises par les actions ou décisions que nous souhaitons effectuer.

Exemple 1.8 (Types d'actions envisagées dans ce cours)**:**

(i) *Estimation ponctuelle* : pour un modèle statistique $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$, on cherche à estimer une grandeur d'intérêt $g(\theta)$, où $\mathrm{P}_\theta \in \mathcal{P}$ est la loi des observations, et où g est une fonction sur l'espace des paramètres à valeur dans $\mathcal{A}$, par exemple, $\mathcal{A} = \mathbb{R}$ ou $\mathcal{A} = \mathbb{R}^d$, $d \geq 1$. Les actions entreprises sont donc des estimations $\gamma \in \mathcal{A}$, qu'on espère être proche de $g(\theta)$. Si la paramétrisation est identifiable, on peut considérer la fonction identité $g(\theta) = \theta$. L'estimation de $g(\theta)$ consiste alors à identifier la loi P_θ à l'origine des observations X.

(ii) *Test statistique* : deux actions peuvent être entreprises : accepter ou rejeter une hypothèse de la forme « $\theta \in \Theta_0$ », où Θ_0 est un sous-ensemble de Θ. Par convention, nous prendrons $\mathcal{A} = \{0, 1\}$, où 1 correspond au rejet de l'hypothèse.

(iii) *Région de confiance* : l'espace des actions $\mathcal{A}$ est composé de sous-ensembles de Θ. Dans ce cas l'objectif est de déterminer un ensemble $\Theta_0 \subset \Theta$ qui contient θ.

(iv) *Prédiction* : L'espace $\mathcal{A}$ est ici beaucoup plus grand. Dans le cas où les observations sont composées d'une variable expliquée $y \in \mathcal{Y}$ et d'une variable explicative $z \in \mathcal{Z}$, alors

$$\mathcal{A} = \{h : h \text{ est une fonction de } \mathcal{Z} \to \mathcal{Y}\},$$

où $h(z)$ représente la prédiction que nous pouvons faire pour y, ayant observé la valeur explicative z.

Comme on le voit, il y a beaucoup de types d'espaces d'actions possibles, assez diverses l'une de l'autre.

Une *règle de décision* est alors définie comme une fonction $\delta : \mathcal{X} \to \mathcal{A}$.

Exemple 1.9 (Règles de décisions associées aux actions de l'exemple 1.8)**:**
En fonction du type d'actions envisagée, la règle de décision δ sera appelée

(i) un *estimateur* (cadre de l'estimation ponctuelle). Un estimateur est alors une statistique (une fonction des observations) à valeurs dans $\mathbb{R}^d$;

(ii) une *procédure de test* (cadre des tests statistiques) : une procédure de test est alors fonction des observations à valeurs dans $\{0, 1\}$

(iii) une *région de confiance* : la procédure de décision est alors un ensemble défini en fonction des observations ;

(iv) un *prédicteur* (cadre de la prédiction) : Un prédicteur est une fonction définie sur $\mathcal{Z}$ qui dépend uniquement des observations.

Une remarque importante : le résultat de la règle de décision appliquée à une observation est de nature aléatoire puisque le modèle d'observation est lui-même aléatoire. Pour comparer des règles de décisions (et, à terme, choisir d'appliquer la « meilleure » règle dans un certain sens), la première étape consiste à comparer les actions. La hiérarchie de préférence entre les actions dépend de la perte encourue, pour une action a et une loi sous-jacente P_θ fixées.

Définition 1.6.1. *Une* fonction de perte *est une application L définie de $\Theta \times \mathcal{A}$ dans $\mathbb{R}_+$ qui permet de hiérarchiser les actions à θ fixé. Ainsi, sous la loi $\mathrm{P}_\theta \in \mathcal{P}$, l'action $a \in \mathcal{A}$ est meilleure que l'action $a' \in \mathcal{A}$ si*

$$L(\theta, a) \leq L(\theta, a').$$

On voit que la hiérarchie entre a et a' découlant d'une fonction de perte L n'est généralement pas absolue : pour une autre valeur θ' du paramètre, il se peut que l'inégalité ci-dessus soit inversée. La fonction de perte est généralement imposée par la nature du problème considéré et les circonstances extérieures à l'analyse (préférences de l'individu, fonctionnement interne d'une entreprise, sensibilité plus ou moins grande d'un individu à tel ou tel traitement,...)

Exemple 1.10 (Fonctions de pertes possibles dans le cadre de l'exemple 1.8)**:**

(i) *Erreur d'estimation ponctuelle* : si g est à valeurs dans $\mathbb{R}$, la fonction de perte la plus couramment utilisée est la perte (ou coût) quadratique

$$L(\theta, \gamma) = (g(\theta) - \gamma)^2 \ .$$

D'autres choix sont bien entendu possibles, mais ils sont en général plus délicats à utiliser. Nous pouvons par exemple considérer l'erreur absolue, $L(\theta, \gamma) = |g(\theta) - \gamma|$, qui pénalise moins les grandes valeurs de l'erreur, ou l'erreur quadratique tronquée, $L(\theta, \gamma) = \min\left((g(\theta) - \gamma)^2, d^2\right)$, qui a un effet similaire. Si $g = (g_1, \ldots, g_d)$ et $\gamma = (\gamma_1, \ldots, \gamma_d)$ sont des vecteurs, des exemples de fonction de coût sont les normes habituels sur les espaces de dimensions finies, par exemple :

$$L(\theta, \gamma) = d^{-1} \sum_{i=1}^{d} (\gamma_j - g_j(\theta))^2,$$

$$L(\theta, \gamma) = d^{-1} \sum_{i=1}^{d} |\gamma_j - g_j(\theta)|,$$

$$\vdots$$

(ii) *Erreur du test* : dans ce cas, on pose $\Theta_1 = \Theta \setminus \Theta_0$ et

$$\begin{aligned} L(\theta, a) &= 0 \quad \text{si } \theta \in \Theta_a \qquad \text{(Décision correcte)} \\ L(\theta, a) &= 1 \quad \text{si } \theta \notin \Theta_a \qquad \text{(Décision erronée).} \end{aligned}$$

(iii) *Erreur de localisation* : on rappelle que l'action a est un sous-ensemble de Θ. Comme dans le cas du test, l'erreur est à valeurs dans $\{0, 1\}$:

$$\begin{aligned} L(\theta, a) &= 0 \quad \text{si } \theta \in a \\ L(\theta, a) &= 1 \quad \text{si } \theta \notin a. \end{aligned}$$

(iv) *Erreur de prédiction* : on reprend le cadre défini au point iv : chaque élément θ de Θ définit une loi P_θ sur $\mathcal{Y} \times \mathcal{Z}$. Soit $h \in \mathcal{A}$, c'est-à-dire une fonction $\mathcal{Z} \to \mathcal{Y}$. On peut par exemple considérer l'erreur quadratique moyenne de prédiction

$$L(\theta, h) = \mathbb{E}_\theta[(Y - h(Z))^2],$$

où (Y, Z) est un couple de loi donné par P_θ.

Une fois fixée une fonction de perte L, on cherche à se donner une bonne "règle de décision" δ. Malheureusement, même à θ fixé, comme l'action $a = \delta(X)$ est aléatoire, la perte encourue $L(\theta, \delta(X))$ l'est aussi. Ceci justifie de considérer une *perte moyenne*, qu'on appellera un *risque*.

Définition 1.6.2. *Soit $\delta : \mathcal{X} \mapsto \mathcal{A}$ une règle de décision. Son risque sous la loi $\mathrm{P}_\theta \in \mathcal{P}$ est défini par*

$$R(\theta, \delta) = \mathbb{E}_\theta \left[L(\theta, \delta(X)) \right] \in \overline{\mathbb{R}}_+ \stackrel{\text{def}}{=} \mathbb{R}_+ \cup \{\infty\} \, . \tag{1.5}$$

Le risque est bien une quantité déterministe, qui ne dépend plus de l'aléatoire X car elle est définie par intégration par rapport x, mais qui dépend du choix de δ et surtout du paramètre θ *inconnu*. Ainsi, il n'est pas possible en général d'ordonner totalement des procédures de décisions, car une "bonne" règle pour un certain paramètre θ_1 (avec un risque $R(\theta_1, \delta)$ faible) peut s'avérer "mauvaise" pour un autre paramètre θ_2. L'exemple 1.11 ci-dessous illustre ce point important. On verra plus tard dans ce cours des critères supplémentaires permettant d' "éliminer θ" dans la définition du risque (risques bayésiens et minimax) et ainsi de choisir une décision "optimale" dans un sens qui reste à préciser.

Exemple 1.11 (Prospection pétrolière, d'après (Bickel, Doksum, 2000))**:**
Nous considérons un modèle statistique paramétrique $(\mathrm{P}_\theta, \theta \in \Theta)$ où l'espace des paramètres Θ est réduit à $\Theta = \{\theta_1, \theta_2\}$. Pour fixer les idées, dans un problème de prospection pétrolière, θ_1 correspond au fait qu'un champ est productif, et θ_2, qu'il ne l'est pas. L'espace des actions $\mathcal{A}$ comporte trois éléments, $\mathcal{A} = \{a_1, a_2, a_3\}$, par exemple nous pouvons forer a_1 pour chercher le pétrole, vendre le champ a_2 à un tiers, ou partager les droits de prospection et d'exploitation a_3. A chaque action est associée une perte dépendant du paramètre :

	Forage	Vente	Partage
	a_1	a_2	a_3
θ_1	0	10	5
θ_2	12	1	6

TABLE 1.1 – Fonction de perte $L(\theta, a)$

Par exemple, s'il y a du pétrole et que nous forons, la perte est 0. S'il n'y a pas de pétrole et que nous forons la perte est 12, et ainsi de suite (voir tableau 1.1). Nous réalisons une expérience pour obtenir une information sur la valeur du paramètre θ. Cette expérience livre une mesure $X \in \mathcal{X} = \{0, 1\}$, et la loi de X est donnée par le tableau des fréquences 1.2.

	$x = 0$	$x = 1$
θ_1	0.3	0.7
θ_2	0.6	0.4

TABLE 1.2 – Fréquence relative

La mesure X représente, par exemple, un type de formation géologique. Des expériences précédentes ont montré que, lorsque le champ est productif (θ_1), on observait des formations de type 0 avec une probabilité 0.3 et des formations de type 1 avec une probabilité 0.7, alors que s'il n'y a pas de pétrole, le formations de type 0 et 1 étaient observées avec des fréquences relatives égales à 0.6 et 0.4. Comme l'ensemble des observations et l'ensemble des actions est fini, il n'y a qu'un nombre fini de décision possibles, qui sont données dans la table 1.3.

	1	2	3	4	5	6	7	8	9
$x=0$	a_1	a_1	a_1	a_2	a_2	a_2	a_3	a_3	a_3
$x=1$	a_1	a_2	a_3	a_1	a_2	a_3	a_1	a_2	a_3

TABLE 1.3 – Fonctions de décision

La procédure δ_1 par exemple consiste à effectuer l'action a_1 indépendamment des résultats de la mesure x. La procédure δ_2 consiste à « faire » a_1 si $x = 0$ et a_2 si $x = 1$ et ainsi de suite. Le risque d'une procédure de décision δ est donnée par

$$R(\theta, \delta) = \mathbb{E}_\theta \left[L(\theta, \delta(X)] = \sum_{i=1}^{3} L(\theta, a_i) \mathrm{P}_\theta [\delta(X) = a_i \right].$$

Nous pouvons comparer différentes procédures de décision en visualisant les points $[R(\theta_1, \delta)\ R(\theta_2, \delta)]$ pour toutes les procédures δ comme représenté sur la figure 1.1. L'ensemble des risques atteignables

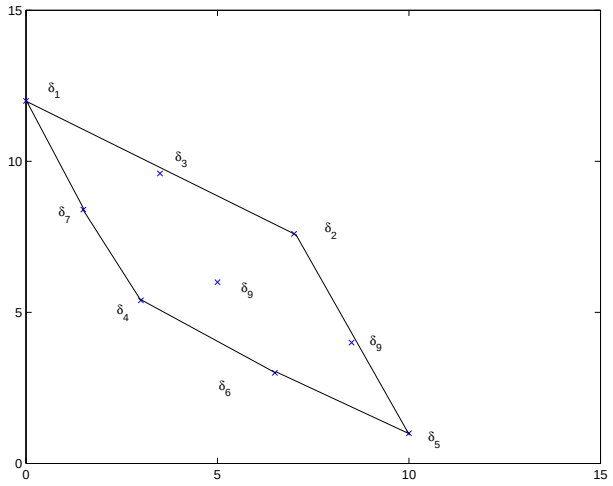


FIGURE 1.1 – Enveloppe convexe de l'ensemble $(R(\theta_1, \delta_i), R(\theta_2, \delta_i)), i = 1, \ldots, 9$

par les règles $(\delta_1, \ldots, \delta_9)$ y sont représentés par les "$\times$".

1.7 Règles randomisées (règles mixtes)*

Dans les parties précédentes, les décisions que nous avons considérées étaient des applications mesurables de l'espace des observations dans l'espace des actions $(\mathcal{A}, \mathcal{B}(\mathcal{A}))$. Comme nous le verrons ci-dessous, il peut être avantageux de considérer une famille de décisions plus générales, appelées *décisions randomisées*. Informellement, l'idée est, une fois observée $x \in \mathcal{X}$, de choisir une action de manière aléatoire selon une distribution qui dépend de x. Par exemple, si l'espace des actions est $\{0, 1\}$, et $\mathcal{X} = \mathbb{R}$, le fait d'observer x puis de lancer un dé et de choisir $a = 1$ si « x est positif et si le résultat du dé est ≥ 5 », ou bien si « x est négatif et le résultat du dé est ≤ 2 » est une règle de décision randomisée.

Plus généralement, une règle de décision randomisée δ^* est une fonction non seulement des données, mais d'une variable aléatoire supplémentaire (le dé dans l'exemple, notée U ci-dessous, et qui est la source de l'aléatoire dans la procédure de décision). Pour simplifier, on considère dans ce cours le cas

où l'espace des actions est fini, $\mathcal{A} = \{1, \dots, K\}$. Plus précisément, une décisions randomisée δ^* est définie par la donnée d'une fonction

$$\Phi : \mathcal{X} \times \mathcal{A} \to [0, 1],$$

telle que $\Phi(x, a)$ soit la probabilité de choisir l'action a en ayant observé x.

Remarque 1.7.1. *La fonction d'ensemble $\Phi(x, \cdot) : A \subset \mathcal{A} \mapsto \Phi(x, A)$ est une loi de probabilité sur $\mathcal{A}$ pour tout x fixé. Si l'on suppose de plus (ce sera toujours le cas en pratique) que l'application $x \mapsto \Phi(x, A)$ est mesurable, quel que soit $A \subset \mathcal{A}$, Φ est appelée* noyau de transition.

La règle de décision δ est alors définie par

$$\delta^* = \Delta(X, U)$$

où U est une variable aléatoire à valeurs dans $\mathcal{U}$ indépendante de X, et où

$$\Delta : \mathcal{X} \times \mathcal{U} \to \mathcal{A}$$

est une fonction mesurable correctement choisie, c'est-à-dire telle que $\mathbb{P}\big[\Delta(x, U) = a\big] = \Phi(x, a)$ pour tout $x \in \mathcal{X}$. On peut par exemple choisir $\mathcal{U} = [0, 1]$, et U une variable aléatoire uniforme sur $[0, 1]$, puis poser, pour $1 \leq j \leq K$,

$$\Delta(x, u) = j \text{ si } u \in \Big[\sum_{i \leq j-1} \Phi(x, i), \sum_{i \leq j} \Phi(x, i) \Big[.$$

Alors on a bien, pour tout $x \in \mathcal{X}$, $\mathbb{P}(\delta^* = j | X = x) = \mathbb{P}\big[\Delta(x, U) = j\big] = \Phi(x, j)$.

On peut facilement construire une règle mixte à partir de règles simples : soient par exemple $\delta^1, \dots, \delta^r$ un ensemble de r règles simples. On peut former une règle randomisée en combinant les δ^j : on se donne un vecteur de poids $p^1, \dots, p^r$, en l'on considère la règle mixte δ^* donnée par le noyau de transition

$$\Phi(x, a) = \sum_{j=1}^{r} p^j \mathbb{1}_{\delta^j(x) = a}$$

("vote" proportionnel des δ^j). On vérifie facilement que le risque d'une telle procédure randomisée s'écrit comme une moyenne pondérée (avec les poids p^j) des risques des procédures non randomisées δ^j. Ainsi l'ensemble des risques atteignables par les règles randomisées formées à partir des règles simple est l'enveloppe convexe des risques atteignables par les règles simple. Dans l'exemple 1.11, c'est le polygone de la figure 1.1.

1.8 Résumé du chapitre

Récapitulons les éléments constitutifs de l'analyse statistique introduits dans ce chapitre

- L'observation X, variable aléatoire définie sur l'espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ à valeurs dans l'espace d'observation $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. la probabilité $\mathbb{P}$ est inconnue, et l'on s'intéresse à la loi P (elle aussi inconnue) de X induite par $\mathbb{P}$: $\mathrm{P}(A) = \mathbb{P}(X \in A)$
- Le modèle statistique $\mathcal{P}$: l'ensemble des "lois candidates" pour l'observation X. C'est une famille de lois de probabilités définies sur l'espace des observations $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. On suppose que $\mathrm{P} \in \mathcal{P}$. En pratique, on indexe les lois du modèle par un paramètre $\theta \in \Theta$, où Θ est l'espace des paramètres. Le modèle s'écrit alors $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$. Ainsi, le paramètre θ est l'inconnue du problème et l'objectif de l'analyse est de tirer de l'information sur θ à partir des observations X.

- L'espace des actions $\mathcal{A}$ (estimation, test, prédiction, intervalle de confiance,...) : c'est l'ensemble des résultats possibles de l'analyse statistique (le délivrable attendu par le commanditaire).
- Une fonction de perte $L : \Theta \times \mathcal{A} \to \mathbb{R}_+$: la quantité $L(\theta, a)$ est la perte encourue lorsque la loi inconnue est P_θ et que l'on entreprend l'action a. Le choix d'une fonction de perte est, dans l'idéal, dicté par la réalité pratique du problème (considérations économiques).
- Les règles de décision (stratégies) : une règle de décision est une fonction $\delta : \mathcal{X} \to \mathcal{A}$ permettant de choisir d'entreprendre telle ou telle action en fonction des données observées.
- Le risque inhérent à une stratégie pour une loi θ donnée, $R(\theta, \delta) = \mathbb{E}_\theta\big[L(\theta, \delta(X))\big]$: c'est l'espérance de la perte encourue étant données une règle de décision δ et une loi P_θ. Le statisticien cherche des fonctions de décision δ telles que "le risque $R(\theta, \delta)$ soit faible" : Attention, à ce stade du cours, puisque le risque dépend de la loi P_θ inconnue, nous n'avons pas encore les outils pour établir une hiérarchie universelle (c'est-à-dire, indépendante de θ) entre deux règles de décision δ et δ', en l'absence d'information supplémentaire sur θ. Ceci sera l'objet du chapitre 4. Avant cela, nous allons nous intéresser au chapitre 3 à une fonction de perte particulière, la perte quadratique, et au risque associé, le risque quadratique.

Chapitre 2

Estimation ponctuelle

Rappelons brièvement le cadre de l'estimation : On considère un modèle $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$ et une quantité d'intérêt (qu'on appellera parfois paramètre) $g(\theta)$ défini pour tout $\theta \in \Theta$ et à valeur dans $\mathcal{A} \subseteq \mathbb{R}^d$. On dispose d'observations $X \sim \mathrm{P}_\theta$, pour un certain $\theta \in \Theta$ inconnu. On cherche à construire un estimateur $\widehat{g}$ pour la quantité $g(\theta)$, c'est-à-dire, rappelons-le, une fonction des observations $\widehat{g} : \mathcal{X} \to \mathcal{A}$. Pour une observation $X = x$, la quantité $\widehat{g}(x)$ sera appelée une *estimation* du paramètre inconnu $g(\theta)$. La fonction $\widehat{g}$ sera appelée *estimateur*.

2.1 M et Z–estimateurs

Une classe importante d'estimateurs consiste à minimiser

$$t \mapsto M(X, t)$$

sur $t \in \mathcal{A}$, où M est une fonction, dite *fonction de contraste* définie sur $\mathcal{X} \times \mathcal{A}$ à valeurs dans $\mathbb{R} \cup \{+\infty\}$. On notera l'ensemble des points qui minimisent $t \mapsto M(X, t)$ par

$$\arg\min_{t \in \mathcal{A}} M(X, t) \stackrel{\text{def}}{=} \left\{ t \in \mathcal{A} : \forall t' \in \mathcal{A}, M(X, t) \leq M(X, t') \right\} .$$

En toute généralité cet ensemble peut être n'importe quel sous–ensemble de $\mathcal{A}$, y compris l'ensemble vide (*non–existence du minimum*). L'existence peut être garantie, par exemple, par des propriétés de continuité de $M(X, t)$ en t et de compacité de $\mathcal{A}$. Quand cet ensemble est un singleton (*unicité du minimum*), on l'identifiera à l'élément qu'il contient, par exemple,

$$\widehat{g} = \arg\min_{t \in \mathcal{A}} M(X, t) \in \mathcal{A} . \tag{2.1}$$

On voit que $\widehat{g}$ est une statistique, et donc un estimateur envisageable pour $g(\theta)$, où P_θ est la loi de X (bien que pour l'instant, rien n'indique que ce soit un "bon" estimateur). Un estimateur de la forme (2.1) s'appelle un M–estimateur.

Parfois, $\widehat{g}$ peut être obtenu en calculant la dérivée (ou les dérivées partielles pour $d \geq 1$) de la fonction $t \mapsto M(X, t)$. Autrement dit, dans "les bons cas", le M–estimateur $\widehat{g}$ peut être défini par

$$\widehat{g} \text{ est solution de } \Psi(X, t) = 0, \quad t \in \mathcal{A} ,$$

où $\Psi(X, \cdot)$ est une fonction de $t \in \mathcal{A}$ à valeurs dans $\mathbb{R}^d$. Par exemple, si $g = (g_1, \ldots, g_d)$, alors $\Psi = (\psi_1, \ldots, \psi_d)$ et la notation ci-dessus est un façon concise d'écrire le système d'équations

$$\begin{cases} \psi_1(X, t) = 0 \\ \quad \vdots \qquad\quad , \quad t \in \mathcal{A}. \\ \psi_d(X, t) = 0 \end{cases}$$

Les systèmes d'équations comme ci-dessus sont appelés *équations d'estimation*. Le système d'équations considéré ci-dessus est obtenu par dérivation d'un contraste mais nous rencontrerons dans la suite des équations d'estimation qui ne dérivent pas d'un contraste (voir le paragraphe 2.3). Dans tous ces cas, un estimateur défini comme solution d'un système d'équations s'appelle un Z–estimateur .

Quand il n'est pas possible d'évaluer exactement le point qui minimise le critère M ou de calculer les solutions du système d'équations Ψ, certaines procédures numériques (algorithmes d'optimisation) peuvent néanmoins garantir que

$$M(X, \widehat{g}) \leq \inf_{t \in \mathcal{A}} M(X, t) + \epsilon \quad \text{ou} \quad \|\Psi(X, \widehat{g})\| \leq \epsilon \,, \tag{2.2}$$

où $\epsilon > 0$ est choisi par l'utilisateur.

Les paragraphes suivants proposent des constructions possibles de M- et Z-estimateurs, sans toutefois répondre à la question :

Comment choisir au mieux M ou Ψ pour estimer $g(\theta)$ sous l'hypothèse que $X \sim \mathrm{P}_\theta$?

Une première réponse possible à cette question sera donnée au chapitre 3. Néanmoins les estimateurs obtenus par les constructions que nous proposerons ne sont

— soit pas toujours explicites (mais plutôt obtenus en pratiques par des procédures numériques qui se contentent de garantir (2.2)),

— soit, quand ils sont explicites, pas toujours sans biais.

Les propriétés asymptotiques des M et Z–estimateurs, qui sortent du cadre de ce cours, apportent une approche alternative pour comparer leur qualité.

2.2 Méthode des moindres carrés

La méthode des moindres carrés est la technique d'estimation de paramètres la plus ancienne. Initialement proposée par Gauss en 1795 pour l'étude du mouvement des planètes, elle fut formalisée par Legendre en 1810. Elle occupe, aujourd'hui encore, une place centrale dans l'arsenal des méthodes d'estimation : son importance pratique est considérable.

Considérons un modèle de régression semi-paramétrique. On observe $X = [X_i = (Y_i, \mathbf{z}_i)]_{1 \leq i \leq n}$ et l'on suppose que

$$Y_i = \varphi(\theta; \mathbf{z}_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

avec

1. $\boldsymbol{\epsilon} = [\epsilon_1 \ \ldots \ \epsilon_n]^T$ vérifiant les hypothèses de Gauss–Markov : $\mathbb{E}[\boldsymbol{\epsilon}] = 0$ et $\mathbb{V}\mathrm{ar}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}$,
2. $\theta \in \mathbb{R}^d$,
3. φ est une fonction de régression ou fonction de lien, supposée connue.

Le modèle de *régression linéaire* correspond au cas où φ est linéaire. Ce modèle est semi-paramétrique dans la mesure où la loi de $\boldsymbol{\epsilon}$ n'est pas entièrement spécifiée.

Dans cet exemple, la grandeur d'intérêt est le paramètre θ lui-même, de sorte que l'on pose $g(\theta) = \theta$, et l'on cherche à construire un estimateur $\widehat{\theta} = \widehat{\theta}(X)$.

Considérons la fonction de contraste définie par

$$M(X,t) = \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \varphi(t;\mathbf{z}_i)\right)^2, \quad t \in \mathbb{R}^d .$$

Nous avons donc

$$\mathbb{E}_\theta[M(X,t)] = \sigma^2 + \frac{1}{n}\sum_{i=1}^{n} \left(\varphi(\theta;\mathbf{z}_i) - \varphi(t;\mathbf{z}_i)\right)^2 .$$

qui est minimum en $t = \theta$ (au moins). Il y a unicité si pour tout t_1 et t_2 dans $\mathbb{R}^d$,

$$\sum_{i=1}^{n} \left(\varphi(t_1;\mathbf{z}_i) - \varphi(t_2;\mathbf{z}_i)\right)^2 = 0 \quad \Rightarrow \quad t_1 = t_2. \tag{2.3}$$

Si, de plus , pour tout $\mathbf{z}$, $t \mapsto \varphi(t;\mathbf{z})$ est continue et si $\lim_{\|t\|\to\infty} \varphi(t;\mathbf{z}) = \infty$ alors, le M–estimateur

$$\widehat{\theta} = \arg\min_{t\in\mathbb{R}^d} M(X,t)$$

est correctement défini (c'est-à-dire, ce minimum existe et est unique) dès que le vecteur d'observation $\mathbf{Y} = [Y_1 \ \ldots \ Y_n]^T$ admet un unique projecteur $\widehat{\mathbf{Y}} \in \mathbb{R}^n$ sur l'ensemble

$$\left\{ [\varphi(t;\mathbf{z}_1) \ \ldots \ \varphi(t;\mathbf{z}_n)]^T \ : \ t \in \mathbb{R}^d \right\} ,$$

ce qui arrive en général presque sûrement. L'estimateur $\widehat{g}$ ainsi défini est appelé l'*estimateur des moindres carrés*. Si la fonction $t \mapsto \varphi(t;\mathbf{z})$ est différentiable sur $\mathbb{R}^d$ pour tout $\mathbf{z}$, l'estimateur des moindres carrés est aussi solution des équations d'estimation

$$\sum_{i=1}^{n} \frac{\partial\varphi}{\partial t_j}(t;\mathbf{z}_i) Y_i = \sum_{i=1}^{n} \frac{\partial\varphi}{\partial t_j}(t;\mathbf{z}_i)\varphi(t,\mathbf{z}_i), \quad 1 \le j \le d.$$

Dans le cas non linéaire, on a recours pour résoudre les équations d'estimations à des procédures numériques, généralement itératives (algorithme de Gauss–Newton par exemple).

2.3 Méthode des moments

La méthode des moments a pour objectif de construire des M ou Z–estimateurs. On se restreint dans ce cours au cas le plus simple d'application de la méthode : celui de l'estimation du paramètre d'une loi. La méthode des moments est alors aussi appelée *principe de substitution* , pour des raisons qui apparaîtront clairement ci-dessous.

On se donne un modèle statistique $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$. On dispose d'un n-échantillon i.i.d. de loi $\mathrm{P} = \mathrm{P}_{\theta_0} \in \mathcal{P}$, c'est-à-dire, $X = (X_1, \ldots, X_n) \in \mathcal{X}^n$ avec $X_1, \ldots, X_n$ indépendants de même loi P_{θ_0} sur $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. On considère le problème de l'estimation de $\theta_0 \in \Theta$. L'espace des actions est donc $\mathcal{A} = \Theta$.

Supposons que l'on dispose de p fonctions $\varphi_1, \ldots, \varphi_p$ définies sur $\mathcal{X}$ à valeurs réelles et intégrables pour tout θ, ($\mathbb{E}_\theta|\varphi_i| < \infty$, pour $i \in \{1, \ldots, p\}$ et $\theta \in \Theta$), telles que l'on puisse

« retrouver θ » dès que l'on connaît la valeur des espérances $\mathbb{E}_\theta \varphi_i(X)$. Cette hypothèse est formalisée ci-dessous. Pour condenser l'écriture, on note $\varphi = (\varphi_1, \ldots, \varphi_p)$, et on introduit la fonction de θ à valeurs dans $\mathbb{R}^p$,

$$\Phi(\theta) = \mathbb{E}_\theta \varphi(X)$$

La fonction Φ est appelée fonction des *moments* associés à φ. Avec ces notations, notre hypothèse s'écrit

(i) [**Injectivité**] Pour tout θ et θ' appartenant à Θ, $\Phi(\theta) = \Phi(\theta')$ implique $\theta = \theta'$,

Appelons $\Phi(\Theta)$ l'image de Θ par Φ. Alors l'hypothèse (i) implique qu'il existe une application réciproque Φ^{-1} définie sur $\Phi(\Theta)$ telle que $\theta = \Phi^{-1}(\Phi(\theta))$ pour tout $\theta \in \Theta$. Pour fixer les idées, prenons l'exemple d'un modèle gaussien, paramétré par $\theta = (\mu, \sigma^2)$ (moyenne et variance). On peut alors prendre $\varphi_i(x) = x^i$, $i = 1, 2$, l'hypothèse (i) ci-dessus est satisfaite, et pour $(m_1, m_2) \in \mathbb{R} \times \mathbb{R}^+$, on a $\Phi^{-1}(m_1, m_2) = (m_1, m_2 - m_1^2)$.

Une idée naturelle consiste à remplacer l'espérance théorique (inconnue) par une version empirique

$$\widehat{\Phi}_n = \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i)$$

(une fonction des données X). En supposant que $\widehat{\Phi}_n \in \Phi(\Theta)$, on prendra comme estimateur $\widehat{\theta}$ la solution de l'équation $\Phi(\theta) = \widehat{\Phi}_n$, c'est à dire $\widehat{\theta} = \Phi^{-1}(\widehat{\Phi}_n)$. L'expression « principe de substitution » vient du fait que l'estimateur $\widehat{\theta}$ est obtenu en *substituant* $\widehat{\Phi}_n$ à Φ dans l'identité $\theta = \Phi^{-1}(\Phi(\theta))$. Par la loi des grands nombres on sait que $\widehat{\Phi}_n \simeq \Phi(\theta_0)$ pour n grand, si θ_0 est le paramètre sous lequel les observations sont générées. Nous ne détaillerons pas dans ce cours la validité asymptotique de la méthode. En pratique, il n'y a pas forcément de solution, mais on peut choisir θ qui minimise l'écart $\|\Phi(\theta) - \widehat{\Phi}_n\|$.

On construit alors le contraste $M_n(t)$ indexé par $t \in \Theta$, défini par

$$M_n(t) = \|\widehat{\Phi}_n - \Phi(t)\|, \qquad t \in \Theta, \tag{2.4}$$

où $\|\cdot\|$ est une norme sur $\mathbb{R}^p$ bien choisie. La statistique $M_n(t)$ s'écrit comme une fonction $M(X, t)$: c'est donc bien un contraste. Pour définir un M–estimateur à partir de ce contraste comme dans (2.1), il faut s'assurer de l'existence et de l'unicité du minimum de la fonction M_n.

Lemme 2.3.1
Sous l'hypothèse (i), s'il existe $\widehat{\theta} \in \Theta$ tel que $M_n(\widehat{\theta}) = 0$, alors $\widehat{\theta}$ est l'unique minimiseur de M_n, c'est-à-dire

$$\widehat{\theta} = \arg\min_{t \in \Theta} M_n(t) \ .$$

Démonstration. Comme la fonction M_n est positive ou nulle, $\widehat{\theta}$ minimise M_n. Le minimum de M_n est donc atteint (en $\widehat{\theta}$). Il suffit maintenant de montrer l'unicité de ce minimum, c'est-à-dire, que si $t \in \Theta$ et $M_n(t) = 0$, alors nécessairement $t = \widehat{\theta}$. Soit donc $t \in \Theta$ tel que $M_n(t) = 0$. Ainsi, $\Phi(t) = \widehat{\Phi}_n = \Phi(\widehat{\theta})$. La condition d'injectivité (i) implique que $t = \widehat{\theta}$. ■

Ainsi, si la fonction Φ est injective et s'il existe $\widehat{\theta}$ tel que $\widehat{\Phi}_n = \Phi(\widehat{\theta})$, on obtient comme estimateur

$$\widehat{\theta} = \Phi^{-1}(\widehat{\Phi}_n) = \arg\min_{t \in \Theta} M_n(t) \ ,$$

quelque soit le choix de la norme $\|\cdot\|$. C'est ce qu'on appelle le *principe de substitution*.

Deux cas particuliers de l'estimation d'un paramètre par la méthode des moments sont donnés ci-dessous.

Exemple 2.1 (temps de survie)**:**
Supposons que les X_i soient des temps de survie modélisés par une loi $\mathrm{P}_\theta = \mathcal{G}amma(\alpha, \lambda)$, de paramètre $\theta = (\alpha, \lambda)$ avec $\alpha > 0$ et $\lambda > 0$ dont la densité est

$$[\lambda^\alpha/\Gamma(\alpha)]x^{\alpha-1}\exp(-\lambda x)\mathbb{1}_{\mathbb{R}_+}(x)\ ,$$

où $\Gamma(\alpha)$ est la fonction Gamma. Construisons un estimateur de $\theta = (\alpha, \lambda) \in \Theta = (0, \infty)^2$. Si on pose $\varphi = (\varphi_1, \varphi_2)$ avec $\varphi_1(x) = x$ et $\varphi_2(x) = x^2$, alors $\Phi(\theta) = \mathbb{E}_\theta(\varphi) = (\mathbb{E}_\theta\varphi_1(X), \mathbb{E}_\theta\varphi_2(X))$ est donnée, pour $\theta = (\theta_1, \theta_2)$, par

$$\mathbb{E}_\theta\varphi_1(X) = \frac{\theta_1}{\theta_2} \quad \text{et} \quad \mathbb{E}_\theta\varphi_2 = \frac{\theta_1(1+\theta_1)}{\theta_2^2}\ ,$$

qui est une fonction inversible de l'ensemble $]0, \infty[^2$ dans lui-même. L'application réciproque Φ^{-1} est

$$\Phi^{-1}(m_1, m_2) = \left(\frac{m_1^2}{m_2 - m_1^2}, \frac{m_1}{m_2 - m_1^2}\right)$$

L'estimateur $\widehat{\theta} = (\widehat{\alpha}, \widehat{\lambda})$ s'écrit alors

$$\widehat{\alpha} = (\bar{X}_n/\widehat{\sigma}_n)^2 \quad \widehat{\lambda} = \bar{X}_n/\widehat{\sigma}_n^2 \quad \text{avec} \quad \bar{X}_n = n^{-1}\sum_{i=1}^n X_i \text{ et } \widehat{\sigma}_n^2 = n^{-1}\sum_{i=1}^n X_i^2 - \bar{X}_n.$$

Il est clair que nous pouvons former de multiples estimateurs de la sorte, en choisissant différents φ_1 et φ_2.

Exemple 2.2 (Modèle de Hardy–Weinberg)**:**
En 1908, un mathématicien anglais, G.H. Hardy, et un médecin allemand W. Weinberg ont formulé une loi connue sous le nom de loi de Hardy–Weinberg. Selon cette loi, les fréquences des allèles d'un gène restent stables de génération en génération dans une population idéale et ne dépendent que des fréquences de la génération initiale (les allèles étant différentes formes d'un gêne). Considérons un gêne à deux allèles (un gêne ayant deux formes différentes). Pour comprendre ce modèle, considérons une population de grande taille. Les individus s'y unissent aléatoirement, impliquant l'union aléatoire des gamètes (chaque gamète étant porteur d'un allèle). Il n'y a pas de migration (aucune copie d'allèle n'est apportée de l'extérieur), pas de mutation, et pas de sélection et les générations sont séparées. Considérons 1 locus à 2 allèles A et a possédant respectivement des fréquences θ et $1-\theta$ à l'équilibre. Quelles vont être les fréquences des différents génotypes AA, Aa et aa ? Pour qu'un individu soit de génotype AA, il faut qu'il ait reçu 1 allèle A de ses 2 parents. Si les gamètes s'unissent au hasard, cet événement se réalise avec la probabilité θ^2. Le raisonnement est identique pour le génotype aa. Enfin, pour le génotype Aa, 2 cas sont possibles : l'individu a reçu A de son père et a de sa mère ou l'inverse, et cet évènement se réalise avec une probabilité $2\theta(1-\theta)$. Les fréquences de Hardy–Weinberg des différents génotypes sont donc données par

Génotype AA	Génotype Aa	Génotype aa
$p_1(\theta) = \theta^2$	$p_2(\theta) = 2\theta(1-\theta)$	$p_3(\theta) = (1-\theta)^2$

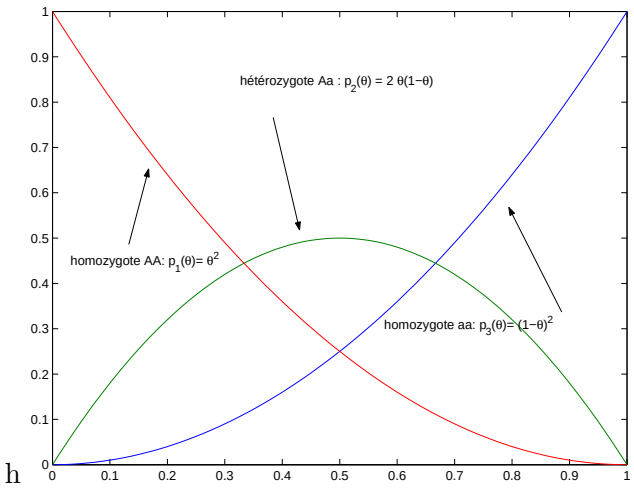


FIGURE 2.1 – Distribution des génotypes suivant le modèle d'équilibre de Hardy–Weinberg pour une population diploïde en fonction de la fréquence de l'allèle A

Considérons une population de n individus. Si N_i est le nombre d'individus dans la population génotypique $i = 1, 2, 3$ correspondant à AA, Aa, aa, alors la variable $\mathbf{N} = (N_1, N_2, N_3)$ suit une loi multinomiale $\mathcal{M}ulti(n, [p_1(\theta), p_2(\theta), p_3(\theta)])$. De façon équivalente, on peut considérer un modèle d'échantillon i.i.d. $X_k \in \mathcal{X} = \{1, 2, 3\}$, $k = 1, \ldots, n$ de loi discrète donnée par $p_i(\theta)$, $i = 1, 2, 3$ et poser

$$N_i = \sum_{k=1}^{n} \mathbb{1}(X_k = i) \ .$$

La statistique N_i/n est le moment empirique associé à la fonction $\varphi_i : x \mapsto \mathbb{1}(x = i)$ définie sur $\mathcal{X}$, et d'espérance

$$\mathbb{E}_\theta \varphi_i = p_i(\theta) \ .$$

Supposons que nous cherchions à estimer la fréquence θ de l'allèle A dans la population. Comme $\theta = \sqrt{p_1(\theta)}$, le principe de substitution nous conduit à l'estimateur $\widehat{\theta} = \sqrt{N_1/n}$. Remarquons aussi que nous avons aussi $\theta = 1 - \sqrt{p_3(\theta)}$, ce qui suggère un autre estimateur $\tilde{\theta} = 1 - \sqrt{N_3/n}$. Parmi 1,705 bébés caucasien nés aux États-Unis en 2000, l'un d'entre eux était porteur de la cystite fibreuse, (homozygote aa). Par suite, $n = 1705$, $N_3 = 1$ et la fréquence de l'allèle A dans la population peut être estimée à

$$\tilde{\theta} = 1 - \sqrt{1/1705} = 0.9758.$$

A partir de cette valeur, on peut estimer la fréquence des homozygotes AA par $\tilde{\theta}^2 = 0.953$ et la fréquence des hétérozygotes (génotype Aa) par $2\tilde{\theta}(1 - \tilde{\theta}) = 0.047$.

Exemple 2.3 (Mélange de deux lois connues)**:**
Soit $(X_1, \ldots, X_n)$ un n-échantillon i.i.d. de loi P_θ, $\theta \in \Theta = (0, 1)$ avec P_θ de densité

$$p(x; \theta) = \theta p_1(x) + (1 - \theta) p_2(x)$$

où $p_1(x)$ et $p_2(x)$ sont deux densités connues définies sur $\mathbb{R}$. Soit φ une fonction de $\mathbb{R}$ dans $\mathbb{R}$. On pose :

$$\mu_j = \int_{\mathbb{R}} \varphi(x) p_j(x) dx \quad \text{pour} \quad j = 1, 2.$$

Alors, l'espérance de $\varphi(X)$ sous P_θ vaut $\mu(\theta) = \theta\mu_1 + (1-\theta)\mu_2$. Avec ces notations, on a :

$$\Phi(t) \stackrel{\text{def}}{=} \mathbb{E}_t \varphi(X) = \mu(t) = t\mu_1 + (1-t)\mu_2 = \mu_2 + t(\mu_1 - \mu_2) \quad t \in [0,1]. \tag{2.5}$$

L'application $\Phi : t \mapsto \mu(t)$ est injective sur $(0,1)$ si et seulement si $\mu_1 \neq \mu_2$ et l'application inverse Φ^{-1} est donnée par

$$\Phi^{-1}(s) = \frac{s - \mu_2}{\mu_1 - \mu_2} \tag{2.6}$$

Par substitution, on en déduit un estimateur de θ :

$$\widehat{\theta} = \Phi^{-1}(\widehat{\Phi}_n) = \frac{n^{-1}\sum_{i=1}^n \varphi(X_i) - \mu_2}{\mu_1 - \mu_2}.$$

Si on choisit, par exemple, de prendre $\varphi(x) = \mathbb{1}(x \leq c)$, alors $\Phi(t) = \mathrm{P}_t(X_1 \leq c)$ est la fonction de répartition de X_1 évaluée c. On en déduit l'expression de l'estimateur de θ :

$$\widehat{\theta} = \frac{\widehat{\Phi}_n(X) - F_2(c)}{F_1(c) - F_2(c)}. \tag{2.7}$$

où $\widehat{\Phi}_n(X) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i \leq c)$ est la proportion d'éléments de l'échantillon inférieurs ou égaux à c et $F_j(c) = \int_{-\infty}^c p_j(x)dx$, $j = 1, 2$. On vérifie aisément que $\widehat{\theta}$ est un estimateur sans biais de θ de variance :

$$\mathrm{EQM}(\widehat{\theta}, \theta) = \frac{F(c)(1-F(c))}{n\,(F_1(c) - F_2(c))^2},$$

où $F(c) = \theta F_1(c) + (1-\theta)F_2(c)$. La variance de cet estimateur dépend du choix du seuil c. On peut résoudre le problème du choix du seuil c en cherchant à la minimiser. Si nous supposons que les densités de probabilité $p_1(x)$ et $p_2(x)$ sont des fonctions continues, la variance est une fonction dérivable du seuil c. Sa dérivée s'annule pour c vérifiant :

$$p(c)[1 - 2F(c)][F_1(c) - F_2(c)] = 2F(c)[1 - F(c)](p_1(c) - p_2(c)) \tag{2.8}$$

On remarque alors que le seuil optimal c dépend de la valeur du paramètre θ *inconnu*. Pour illustrer ces résultats, considérons le mélange de deux lois gaussiennes, p_1 et p_2 densités des lois $\mathcal{N}(\mu_1, \sigma_1^2)$ et $\mathcal{N}(\mu_2, \sigma_2^2)$, dans les deux cas suivants :

	(μ_1, σ_1^2)	(μ_2, σ_2^2)
cas I	(0, 1)	(3,1)
cas II	(0,1)	(1,4)

Dans le cas I, les deux composantes gaussiennes sont bien séparées tandis que dans le cas II elles le sont mal. On peut donc s'attendre à ce que l'estimation de la proportion dans I soit plus aisée que dans II. C'est ce que montre les courbes de la figure 2.2, où nous avons représenté la variance de l'estimateur de la proportion du mélange, en fonction du choix du seuil c, lorsque $\theta = 1/2$. Dans le cas I, on observe que le minimum est atteint pour $c \simeq 1.5$, ce qui n'est pas surprenant car le point d'intersection entre $p_1(x)$ et $p_2(x)$ est en $c = 1.5$, ce qui annule le membre de droite de (2.8) et $F(c) = 1/2$ (c est la médiane de la loi d'observation), ce qui annule le membre de gauche de (2.8), et par conséquent $c = 1.5$ est solution de l'équation (2.8). Dans le cas II, la valeur optimale de c est voisine de 1.

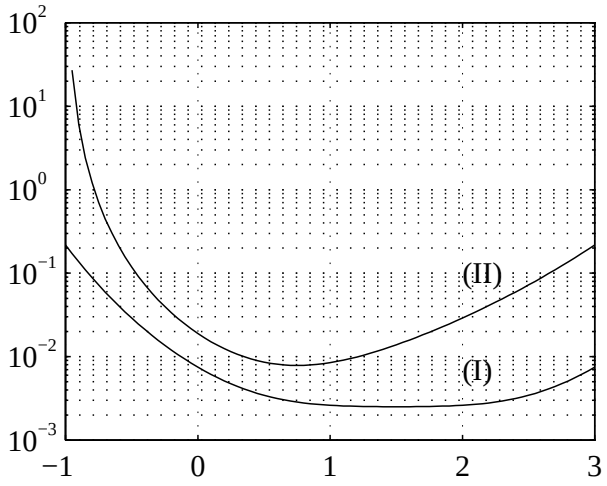


FIGURE 2.2 – Variance de l'estimateur de la proportion d'un mélange de 2 lois gaussiennes en fonction du seuil c, pour $\theta = 1/2$.

2.4 Méthode du Maximum de vraisemblance

La méthode du maximum de vraisemblance a été introduite, dans le cas de modèles d'observation discrets par Gauss en 1821. Toutefois, cette approche est habituellement associée au nom du statisticien anglais Fisher, qui a redécouvert cette méthode d'inférence et a été le premier à donner les bases d'une théorie de l'estimation paramétrique fondée sur la vraisemblance.

Nous nous plaçons dans le cadre d'un modèle dominé indexé par un paramètre $\theta \in \Theta$, $\mathcal{P} = \{\mathrm{P}_\theta,\, \theta \in \Theta\}$ défini sur l'espace d'observation $\mathcal{X}$. On notera $p(\cdot;\theta)$ la densité de P_θ par rapport à la mesure dominante qu'il est inutile de préciser pour la suite. On observe X de loi P_θ. Rappelons que la fonction de vraisemblance, à $X = x$ fixé est la fonction de t, $t \to p(x;t)$.

On appelle *estimateur du maximum de vraisemblance* de θ, l'estimateur associé au contraste $-p(X,\cdot)$, c'est-à-dire tout estimateur $\widehat{\theta}$ vérifiant

$$p(X;\widehat{\theta}) \geq \sup\{p(X;t) \ : \ t \in \Theta\} \ . \tag{2.9}$$

Exemple 2.4 (Nombre moyen d'arrivées dans une file d'attente)**:**
Considérons tout d'abord le modèle discret d'une file d'attente à un serveur. On suppose que le nombre de clients qui arrivent pendant un intervalle de durée fixe suit une loi de Poisson de moyenne $\theta > 0$ et que les nombres observés dans des intervalles disjoints sont des variables aléatoires indépendantes. On effectue une observation $X = (X_1, \cdots, X_n)$ dans n intervalles disjoints de même taille. L'hypothèse d'indépendance implique que la densité par rapport à la mesure de comptage s'écrit :

$$p(x;\theta) = \mathrm{P}_\theta(X_1 = x_1, \cdots, X_n = x_n) = \frac{\theta^{x_1} \cdots \theta^{x_n}}{x_1! \cdots x_n!} \exp(-n\theta)$$

où $x = (x_1, \ldots, x_n)$ est un vecteur d'entiers naturels positifs. En passant au logarithme qui est

une fonction monotone croissante, on obtient la *log–vraisemblance* :

$$\log\left(p(X;t)\right) = t\sum_{i=1}^{n} X_i \; - nt - \sum_{i=1}^{n} \log X_i!$$

En annulant la dérivée par rapport t, on obtient l'estimateur du maximum de vraisemblance

$$\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \; ,$$

qui n'est autre, dans ce cas, que la moyenne empirique des observations.

Exemple 2.5 (Estimation du paramètre inconnu d'une loi uniforme)**:**
Considérons un échantillon i.i.d. $(X_1, \cdots, X_n)$ de v.a. de loi uniforme sur un intervalle $[0, \theta]$, où $\theta > 0$ est le paramètre inconnu à estimer. La densité de ce modèle est

$$p(x;\theta) = \theta^{-1}\mathbb{1}(0 \leq x \leq \theta) \; .$$

La vraisemblance des observations est donnée par

$$p(X_1, \ldots, X_n; \theta) = \begin{cases} 0 & \theta \leq \max(X_1, \ldots, X_n), \\ \theta^{-n} & \theta \geq \max(X_1, \ldots, X_n) \; . \end{cases}$$

et l'estimateur du maximum de vraisemblance est donné par $\widehat{\theta}_n = \max(X_1, \cdots, X_n)$.

Dans le cas d'un échantillon i.i.d., $X = (X_1, \ldots, X_n)$, de densité $p(\cdot;\theta)$, $\theta \in \Theta$, il est pratique de considérer la *log–vraisemblance* définie comme le logarithme de la fonction de vraisemblance

$$L(x;t) = \log \prod_{i=1}^{n} p(x_i;t) = \sum_{i=1}^{n} \log p(x_i;t), \quad x \in \mathcal{X}, t \in \Theta \; . \tag{2.10}$$

Pour une valeur de t fixée, le contraste défini comme l'opposé de la log–vraisemblance $-L(X,t)$ est alors une somme de variables aléatoires réelles i.i.d., ce qui sera utile dans l'analyse de ses propriétés asymptotiques.

Exemple 2.6 (Modèle paramétrique)**:**
Supposons le modèle considéré paramétrique, $\Theta \subseteq \mathbb{R}^p$. Si la vraisemblance $p(x;t)$ est différentiable en t et si l'estimateur du maximum de la vraisemblance $\widehat{\theta}$ est un point intérieur de Θ, alors l'estimateur du maximum de vraisemblance $\widehat{\theta}$ est une solution des *équations de vraisemblance*

$$\nabla_t \log p(x;t) = 0, \quad t \in \mathrm{int}(\Theta) \; . \tag{2.11}$$

C'est donc un Z–estimateur. Il faut faire attention quand les équations de vraisemblance sont utilisées de bien vérifier que la solution trouvée correspond au maximum global de $t \mapsto p(X;t)$ sur $t \in \Theta$ et non un minimum ou un maximum local.

Exemple 2.7 (Echantillon multinomial)**:**
On considère une expérience consistant à tirer indépendamment n éléments dans une population comportant k composantes. Notons $\theta_j = \mathrm{P}_\theta[X_i = j]$ la probabilité de tirer la composante j,

$\sum_{j=1}^{k} \theta_j = 1$, et $N_j = \sum_{i=}^{n} I[X_i = j]$ le nombre d'observations dans la j-ième catégorie. La log–vraisemblance des observations est donnée par

$$\mathbf{t} = (t_1, \ldots, t_k) \to L(X, \mathbf{t}) = \sum_{j=1}^{k} N_j \log(t_j), \quad \mathbf{t} \in \Theta = \left\{ \mathbf{t} : t_j \geq 0, \sum_{j=1}^{k} t_j = 1 \right\}.$$

Pour obtenir l'estimateur du maximum de vraisemblance, nous maximisons la log–vraisemblance par rapport aux $k-1$ paramètres $t_1, \ldots, t_{k-1}$ en posant $t_k = 1 - \sum_{i=1}^{k-1} t_i$. Nous considérons d'abord le cas où tous les N_j sont strictement positifs. Nous avons $L(\mathbf{t}, X) = -\infty$ si l'un des t_j est nul, et donc le maximum de vraisemblance est à l'intérieur du domaine et doit satisfaire les équations de vraisemblance

$$\frac{\partial}{\partial t_j} L(\mathbf{t}, X) = \frac{N_j}{t_i} - \frac{N_k}{t_k} = 0.$$

Par conséquent $\widehat{\theta}_j / \widehat{\theta}_k = N_j / N_k$ et donc $\widehat{\boldsymbol{\theta}} = n^{-1}[N_1 \ \ldots \ N_k]^T$. Le calcul de la dérivée seconde montre que ce point est bien un maximum. Si maintenant il y a des indices j pour lesquels N_j est nul, notons I l'ensemble des indices restant (au moins 1 puisque $n \geq 1$). On se ramène donc à maximiser $\sum_{j \in I} N_j \log(t_j)$ sous les contraintes $t_j \geq 0$ pour $j \in I$ et $\sum_{j \in I} t_j \leq 1$. Il est clair que le maximum sera atteint uniquement si $\sum_{j \in I} t_j = 1$, ce qui donne finalement, en appliquant le résultat précédent $\widehat{\boldsymbol{\theta}} = n^{-1}[N_1 \ \ldots \ N_k]^T$.

Exemple 2.8 (Modèle de Hardy–Weinberg, cf. exemple 2.2)**:**
La vraisemblance en l'observation $\mathbf{N} = [N_1, N_2, N_3]$ est ici donnée par

$$L(\mathbf{N}; t) \propto t^{2N_1} (2t(1-t))^{N_2} (1-t)^{2N_3} \propto t^{2N_1 + N_2} (1-t)^{N_2 + 2N_3}.$$

Si $2N_1 + N_2 > 0$ et $N_2 + 2N_3 > 0$ alors l'estimateur du maximum de vraisemblance est donné par

$$\widehat{\theta} = \frac{2N_1 + N_2}{2n}.$$

Si $2N_1 + N_2 = 0$ ($N_1 = 0$ et $N_2 = 0$), alors la vraisemblance est égale à $(1-t)^n$ qui est maximisée en $\widehat{\theta} = 0$. De façon similaire, si $N_2 = 0$ et $N_3 = 0$, alors $\widehat{\theta} = 1$.

Exemple 2.9 (Echantillon Gaussien)**:**
Soit $(X_1, \cdots, X_n)$ un n-échantillon $\mathcal{N}(\mu, \sigma^2)$. On note $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}^+$. La log–vraisemblance a donc pour expression, pour tout $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, $\mathbf{t} = (t_1, t_2) \in \Theta$,

$$\log p(x; \mathbf{t}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log t_2 - \frac{1}{2t_2} \sum_{i=1}^{n} (x_i - t_1)^2 \,. \tag{2.12}$$

L'estimateur du maximum de vraisemblance $\widehat{\theta}_n = (\widehat{\mu}_n, \widehat{\sigma}_n^2)$ s'obtient en résolvant les équations de log–vraisemblance :

$$\frac{\partial p}{\partial t_1}(X; \mathbf{t}) = 0 \quad \text{et} \quad \frac{\partial p}{\partial t_2}(X; \mathbf{t}) = 0, \quad \mathbf{t} \in \Theta \,.$$

Ces équations de vraisemblance ont une solution unique dans Θ et on obtient :

$$\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{et} \quad \widehat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \widehat{\mu}_n)^2 \tag{2.13}$$

Exemple 2.10 (Loi double-exponentielle translatée)**:**
Considérons le modèle donné par la famille densités $\{p_\theta(x) = \frac{1}{2}\exp(-|x-\theta|),\, \theta \in \mathbb{R}\}$ (la famille des lois double-exponentielles avec la mesure de Lebesgue pour mesure dominante). Pour ce modèle, dans le cas d'un n-échantillon i.i.d. $X = (X_1, \ldots, X_n)$, la log–vraisemblance s'écrit

$$L(X,t) = -\sum_{k=1}^{n} |X_i - t|, \quad t \in \mathbb{R}\,.$$

Elle est maximum en tout point t tel que

$$\#\{i:\, X_i \le t\} = \#\{i:\, X_i \ge t\}\,.$$

Il existe au moins un tel point t et, comme les X_i sont distincts p.s., il y a unicité uniquement si n est impair, auquel cas l'estimateur $\widehat{\theta}_n$ ainsi défini est la *médiane empirique*.

Exemple 2.11 (Loi uniforme sur un intervalle quelconque)**:**
Soient $\{X_n\}_{n\ge 0}$ des observations i.i.d distribuées suivant une loi uniforme sur l'intervalle $[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$, $\theta \in \mathbb{R}$. La fonction de vraisemblance $\theta \to p(x_1, \cdots, x_n; \theta)$ est donnée pour tout $x = (x_1, \ldots, x_n)$ et $t \in \mathbb{R}$ par :

$$p(x;t) = \begin{cases} 1 & \theta \in [M_n(X) - 1/2, m_n(X) + 1/2] \\ 0 & \text{sinon} \end{cases}$$

où $M_n(X) = \max(x_1, \cdots, x_n)$ et $m_n(X) = \min(x_1, \cdots, x_n)$. La vraisemblance est constante sur l'intervalle $[M_n(X) - 1/2, m_n(X) + 1/2]$ et toute valeur prise dans cet intervalle est un estimateur du maximum de vraisemblance. Considérons par exemple les deux estimateurs suivants :

$$\widehat{\theta}^{(1)} = M_n(X) - \frac{1}{2}, \quad \text{et} \quad \widehat{\theta}^{(2)} = m_n(X) + \frac{1}{2}.$$

On peut établir que :

$$\mathbb{E}_\theta[(\widehat{\theta}_n^{(1)} - \theta)^2] = \mathbb{E}_\theta[(\widehat{\theta}_n^{(2)} - \theta)^2] = \frac{2}{(n+1)(n+2)} \tag{2.14}$$

Remarquons toutefois tous les estimateurs du maximum de vraisemblance n'ont pas le même risque quadratique (cf. Chapitre 3). En particulier l'estimateur $\widehat{\theta}^{(3)} = (M_n(X) + m_n(X))/2$ vérifie

$$\mathbb{E}_\theta\left[(\widehat{\theta}_n^{(3)} - \theta)^2\right] = \frac{1}{2(n+1)(n+2)}\,,$$

et a donc un risque quadratique plus faible que $\widehat{\theta}^{(1)}$ et $\widehat{\theta}^{(2)}$.

2.5 Famille exponentielle*

De nombreux modèles « classiques » se prêtent bien à l'estimation par maximum de vraisemblance, en particuliers les modèles de type « famille exponentielle », définis comme suit.

Définition 2.5.1 (famille exponentielle)**.** *Un modèle* $\{\mathrm{P}_\theta, \theta \in \Theta \subset \mathbb{R}^d\}$ *sur* $\mathcal{X}$ *est appelé* famille exponentielle *de dimension* d*, si le modèle est dominé par une mesure* μ *sur* $\mathcal{X}$*, par rapport à laquelle* P_θ *admet une densité* p_θ *de la forme*

$$p(x;\theta) = h(x)\exp\left[\langle \eta(\theta), T(x)\rangle - B(\theta)\right], \quad x \in \mathcal{X}, \theta \in \Theta, \tag{2.15}$$

où $T : \mathcal{X} \to \mathbb{R}^p$ est une statistique multivariée, $\eta : \Theta \to \mathbb{R}^p$ est une fonction du paramètre, appelée « paramètre naturel » et $e^{-B(\theta)}$ est une constante de normalisation assurant que l'intégrale $\int_{\mathcal{X}} p_\theta(x)\mathrm{d}\mu(x) = 1$.

Dans ce cas, on appelle « espace des paramètres naturels » du modèle l'ensemble

$$\mathcal{E} = \{\eta \in \mathbb{R}^p : Z(\eta) := \int_{\mathcal{X}} h(x)e^{\langle \eta, T(x)\rangle}\mu(\mathrm{d}x) < \infty\}.$$

Remarque 2.5.2. *D'après la définition, la constante de normalisation est choisie de sorte que $Z(\eta) = e^{B(\theta)}$. Ainsi, la constante ne dépend de θ qu'à travers $\eta(\theta)$ et on peut toujours choisir la « paramétrisation naturelle » $\mathcal{P} = \{\mathrm{P}_\eta : \eta \in \mathcal{E}\}$ où P_η admet pour densité*

$$p(x;\eta) = h(x)e^{\langle \eta, T(x)\rangle - A(\eta)}, \quad x \in \mathcal{X}, \eta \in \mathcal{E}, \tag{2.16}$$

avec $A(\eta) = -\log \int_{\mathcal{X}} h(x)e^{\langle \eta, T(x)\rangle}\mu(\mathrm{d}x) = -\log Z(\eta)$.

La définition d'une famille exponentielle, bien que d'apparence restrictive, recouvre un grand nombre d'exemples classiques : par exemple, on vérifie facilement (exercice) que les modèles de Bernoulli, binomial, de Poisson, le modèle exponentiel donné par $p(x,\theta) = \theta e^{-\theta x}, x \geq 0, \theta > 0$, le modèle gaussien (avec $p = 2$ et $\eta(\mu,\sigma^2) = (\mu/\sigma^2, -1/2\sigma^2)$) sont des modèles exponentiels.

2.6 Maximum de vraisemblance pour la famille exponentielle*

Les questions d'existence et d'unicité de l'estimateur du maximum de vraisemblance peuvent être traitées de façon assez élégantes et complètes dans le cas de la famille exponentielle canonique. Ceci découle assez directement de la concavité stricte de la log–vraisemblance en fonction du paramètre canonique η. Soit $(X_1, \ldots, X_n)$ un n-échantillon d'une expérience statistique $(\mathrm{P}_\eta, \eta \in \mathcal{E})$ où P_η est une famille exponentielle d-dimensionnelle de densité :

$$p_\eta(x) = h(x)\exp(\langle \eta, T(x)\rangle - A(\eta)), \quad \eta \in \mathcal{E},$$

par rapport à une mesure de domination μ. $\mathcal{E}$ est l'espace des paramètres canoniques,

$$\mathcal{E} = \left\{\eta \in \mathbb{R}^d, A(\theta) = \log\left(\int h(x)\exp(\langle \eta, T(x)\rangle)\mu(dx)\right) < \infty\right\}.$$

L'ensemble $\mathcal{E}$ est convexe. Nous supposons dans la suite que la famille est régulière, auquel cas cet ensemble est ouvert. Dans la suite, nous supposons toujours que le vrai paramètre $\eta_0 \in \mathcal{E}$. Nous admettons le lemme suivant, qui est une conséquence de la proposition 3.2.11 :

Lemme 2.6.1
La fonction $\eta \to \int h(x)\exp(\langle \eta, T(x)\rangle)\mu(dx)$ est analytique sur Θ et

$$\frac{\partial^p \int h(x)\exp(\langle \eta, T(x)\rangle)\mu(dx)}{\partial\eta_1^{i_1}\cdot\partial\eta_k^{i_k}} = \int h(x)T_1(x)^{i_1}\cdots T_k(x)^{i_k}\exp(\langle \eta, T(x)\rangle)\mu(dx),$$

pour tout entier naturel p et tout $i_1 + \cdots + i_k = p$. Autrement dit,

$$\frac{\partial^p \exp[A(\eta)]}{\partial\eta_1^{i_1}\cdot\partial\eta_k^{i_k}} = \mathbb{E}[T_1(X)^{i_1}\cdots T_k(X)^{i_k}]\exp[A(\eta)]$$

Ce lemme implique que la log–vraisemblance

$$L(\eta, x) = \log p(x;\eta) = \text{Constante}(x) + \langle T(x), \eta\rangle - A(\eta)$$

est (indéfiniment) différentiable sur $\mathcal{E}$ par rapport à η. En particulier, pour $p = 1, 2$ on obtient

$$\nabla_\eta A(\eta) = \mathbb{E}_\eta[T(X)]\ , \quad \nabla^2_\eta A(\eta) \stackrel{\text{def}}{=} \left(\frac{\partial^2 A(\eta)}{\partial\eta_i \partial\eta_j}\right)_{i,j\leq d} = \mathbb{C}\text{ov}_\eta(T(X))$$

(où l'on note $\mathbb{C}\text{ov}(Y)$ la matrice de variance -covariance d'un vecteur aléatoire Y). Ainsi, la matrice hessienne de la log–vraisemblance est $\left(\frac{\partial^2[A(\eta)+\langle T(x),\eta\rangle+\text{Constante}]}{\partial\eta_i\partial\eta_j}\right)_{i,j\leq d} = -\mathbb{C}\text{ov}_\eta(T(X))$, qui est une matrice définie *négative* d'après les propriétés des matrices de variance-covariance. La log-vraisemblance est donc une fonction concave. De plus, les dérivées partielles de la log–vraisemblance (ou fonction score) sont données par :

$$\nabla_\eta L(\eta, x) = T(x) - \nabla_\eta A(\eta) = T(x) - \mathbb{E}_\eta[T(X)].$$

Cette relation illustre une propriété qui sera mise en évidence au Chapitre 3 pour les modèles statistiques réguliers : l'espérance du score est nulle. Pour un n-échantillon i.i.d de loi $p_\theta(x)$, les équations de vraisemblance se réduisent ici à :

$$n^{-1}\sum_{i=1}^{n} T(X_i) = \mathbb{E}_\eta[T(X)], \tag{2.17}$$

et les estimateurs du maximum de vraisemblance sont, pour cette famille de loi, des estimateurs obtenus par la méthode des moments en prenant comme fonction des moments $\phi(x) = T(x)$. En supposant que la fonction $\eta \to \mathbb{E}_\eta[T(X)]$ est bijective sur $\mathcal{E}$, la solution de (2.17), *si elle existe*, est unique. La log–vraisemblance étant concave, ce point est nécessairement un maximum. Dans de nombreux cas, nous considérons des familles exponentielles de la forme

$$p(x;\theta) = h(x)\exp(\langle q(\theta), T(x)\rangle - B(\theta)), \quad \theta \in \Theta,$$

où θ est le paramètre. Si la fonction q est une fonction bijective de $\mathbb{R}^d \mapsto \mathbb{R}^d$, et si l'estimateur du maximum de vraisemblance existe pour le paramètre canonique $\widehat{\eta}$, alors on peut vérifier que que $\widehat{\theta} = q(\widehat{\eta})$ est un estimateur de maximum de vraisemblance.

Chapitre 3

Risque quadratique

3.1 Risque quadratique

Considérons un modèle statistique $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$ défini pour un espace d'observations $\mathcal{X}$. On s'intéresse à un paramètre réel d'intérêt $g(\theta)$, par exemple l'espérance de X sous P_θ ou un quantile, etc... des exemples ont été donnés au chapitre précédent. On considère dans ce chapitre le problème de l'estimation de $g(\theta)$. On est dans le cadre de l'*estimation ponctuelle* introduit dans les exemples de la section 1.6. L'espace des actions est donc la droite réelle, et une fonction de décision est un estimateur, c'est-à-dire une fonction $\widehat{g} : \mathcal{X} \to \mathbb{R}$. Attention, le paramètre d'intérêt g est une fonction de θ (inconnu), alors que l'estimateur $\widehat{g}$ est une fonction de X (observé).

La fonction de perte la plus courante pour l'estimation d'un paramètre réel est la fonction de perte quadratique, définie pour un paramètre θ et une estimation $\gamma \in \mathbb{R}$ destinée à approcher $g(\theta)$, par

$$L(\theta, \gamma) = (g(\theta) - \gamma)^2.$$

Soit $\widehat{g} : \mathcal{X} \to \mathbb{R}$ un estimateur du paramètre $g(\theta)$. Le risque correspondant est appelé *erreur quadratique moyenne* (EQM) ou *risque quadratique*. Il est donné par

$$\mathrm{EQM}(\theta, \widehat{g}) \stackrel{\text{def}}{=} R(\theta, \widehat{g}) = \mathbb{E}_\theta[(g(\theta) - \widehat{g}(X))^2] \,. \tag{3.1}$$

L'erreur quadratique moyenne dépend de la variance de l'estimateur et de son *biais* défini par :

$$b(\theta, \widehat{g}) \stackrel{\text{def}}{=} \mathbb{E}_\theta[\widehat{g}(X) - g(\theta)]. \tag{3.2}$$

Proposition 3.1.1 (Décomposition biais–variance)
Pour tout $\theta \in \Theta$, l'erreur quadratique moyenne définie en (3.1) se décompose en

$$\mathrm{EQM}(\theta, \widehat{g}) = (b(\theta, \widehat{g}))^2 + \mathbb{V}\mathrm{ar}_\theta(\widehat{g}(X)), \tag{3.3}$$

où $\mathbb{V}\mathrm{ar}_\theta(\widehat{g}(X)) \stackrel{\text{def}}{=} \mathbb{E}_\theta[(\widehat{g}(X) - \mathbb{E}_\theta\widehat{g}(X))^2]$ est la variance de $\widehat{g}(X)$ et $b(\theta, \widehat{g})$ est le biais de l'estimateur $\widehat{g}$.

Démonstration. On écrit, dans le membre de droite de (3.1),

$$\widehat{g}(X) - g(\theta) = [\widehat{g}(X) - \mathbb{E}_\theta\widehat{g}(X)] + [\mathbb{E}_\theta\widehat{g}(X) - g(\theta)] \,.$$

On obtient le résultat en développant le carré et en calculant l'espérance sous P_θ. ■

Exemple 3.1 (Estimation du paramètre de translation)**:**
Soit $X = (X_1, \ldots, X_n)$ un n-échantillon i.i.d. de loi $\mathcal{N}(\mu, \sigma^2)$ (loi gaussienne de moyenne μ et de variance σ^2). Nous utilisons comme estimateur de μ la moyenne empirique, $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ et nous utilisons une fonction de coût quadratique. Le biais de cet estimateur est nul et sa variance est égale à :

$$\mathbb{V}\mathrm{ar}_{\mu,\sigma^2}(\bar{X}) = n^{-2} \sum_{i=1}^n \mathbb{V}\mathrm{ar}_{\mu,\sigma^2}(X_i) = \sigma^2/n.$$

Par conséquent, le risque quadratique est donné par

$$\mathrm{EQM}(\mu, \sigma^2; \bar{X}) = \sigma^2/n,$$

qui ne dépend pas de μ.

Dans le cas où la qualité des mesures (plus précisément la variance σ^2 des erreurs) est connue, l'expression exacte du risque ci-dessus permet de déterminer à l'avance le nombre de mesures nécessaires pour avoir un risque inférieur à un niveau donné $\epsilon > 0$, $n_0 = \lceil \sigma^2/\epsilon \rceil$.

Si σ^2 est inconnue, une évaluation aussi précise du risque est impossible. On peut toutefois estimer la variance des erreurs, par exemple en prenant $\widehat{\sigma}^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$, qui est un estimateur sans biais de σ^2. L'estimation correspondante du risque erreur est cette fois elle aussi sujette à des fluctuations aléatoires et doit être utilisée avec certaines précautions que nous ne détaillerons pas ici.

La décomposition biais-variance et l'exemple ci-dessus suggèrent une des raisons pour lesquelles on utilise le plus souvent le risque quadratique : la simplicité des calculs. L'exemple ci-dessous illustre les difficultés rencontrées avec une autre fonction de perte d'apparence pourtant simple

Exemple 3.2 (perte absolue)**:**
Supposons qu'à la place de la perte quadratique, nous ayons choisi d'utiliser la perte absolue $L(\theta, \gamma) = |g(\theta) - \gamma|$. Comme P_θ ne dépend que de (μ, σ^2) dans le modèle Gaussien, peut alors écrire le risque en fonction de μ et σ^2,

$$R(\mu, \sigma^2; \bar{X}) = \mathbb{E}_{\mu,\sigma^2} |\bar{X} - \mu| \ .$$

Pour le modèle considéré, $\bar{X} - \mu$ suit une loi $\mathcal{N}(0, \sigma^2/n)$, d'où, par un calcul plus délicat,

$$R(\mu, \sigma^2; \bar{X}) = \frac{\sigma}{\sqrt{n}} \int_{-\infty}^{\infty} |t| \ \mathrm{e}^{-t^2/2} \, dt = \frac{\sigma\sqrt{2}}{\sqrt{n\pi}}$$

Si nous ne supposons plus que la distribution des erreurs est gaussienne, mais suit une loi de densité quelconque symétrique f, alors le risque quadratique de l'estimateur de la moyenne est encore donné par $\mathbb{V}\mathrm{ar}_f(X)/n$, mais on ne dispose plus alors d'expression explicite pour le risque absolu (aussi appelé « risque uniforme »), sauf exception : pour évaluer le risque, on a alors recours soit à des méthodes d'intégration numérique, soit à des méthodes de simulation.

En fait, les difficultés apparaissent aussi pour le risque quadratique, lorsque nous utilisons un estimateur autre que la moyenne empirique. Si par exemple nous considérons comme estimateur la médiane, il n'est plus possible de calculer explicitement le risque quadratique, même lorsque la loi est gaussienne, et l'on doit avoir recours à des méthodes numériques.

La décomposition biais–variance (3.3) apporte une simplification dans l'analyse du risque quadratique pour les estimateurs dits *sans biais*.

Définition 3.1.2 (Estimation sans biais de variance minimale)**.** *On dira qu'un estimateur* $\widehat{g}$ *est* sans biais *si*

$$b(\theta, \widehat{g}) = 0 \quad \textit{pour tout} \quad \theta \in \Theta \, .$$

On appelle « classe des estimateurs sans biais » l'ensemble Γ *des estimateurs* $\widehat{g}$ *qui vérifient cette contrainte. Quand il existe, l'estimateur de cette classe qui vérifie*

$$\mathrm{EQM}(\theta, \widehat{g}) \leq \mathrm{EQM}(\theta, \widehat{g}') \quad \textit{pour tout} \quad \theta \in \Theta \quad \textit{et tout} \quad \widehat{g}' \in \Gamma,$$

c'est-à-dire

$$\widehat{g} \in \operatorname*{argmin}_{\widehat{g}' \in \Gamma} \mathrm{EQM}(\theta, \widehat{g}'), \ \textit{pour tout } \theta \in \Theta$$

est appelé estimateur uniformément de variance minimale dans la classe des estimateurs sans biais *(U.V.M.B.).*

3.2 Information de Fisher, Borne de Cramér-Rao

Nous allons dans cette partie établir des bornes inférieures sur le risque quadratique des estimateurs sans biais. Cette borne permet d'évaluer l'écart entre l'estimateur utilisé et une borne ultime, qui n'est pas nécessairement atteinte.

3.2.1 Modèle statistique régulier, information de Fisher

L'information de Fisher est une notion centrale en statistique paramétrique. En un mot, cette quantité représente (comme son nom l'indique et pour des raisons que nous ne détaillerons pas dans ce cours) la quantité moyenne d'information apportée par une observation. Elle est définie sous certaines conditions techniques sur le modèle statistique en jeu, détaillées ci-dessous.

Définition 3.2.1 (Modèle régulier)**.** *Soit* $\mathcal{P} = (\mathrm{P}_\theta, \theta \in \Theta \subset \mathbb{R}^d)$ *un modèle paramétrique dominé par une mesure* μ *:* $\mathrm{P}_\theta(dx) = p_\theta(x)\mu(dx)$*. Nous noterons* $p(x;\theta) \equiv p_\theta(x)$ *la densité. Le modèle est appelé* régulier *si les conditions suivantes sont vérifiées :*

(1) ***Espace des paramètres régulier et support constant :*** *L'espace des paramètres* Θ *est un sous-ensemble ouvert de* $\mathbb{R}^d$ *; et l'ensemble* $\mathfrak{S} = \{x \in \mathbb{R}^q : p(x;\theta) > 0\}$ *ne dépend pas de* θ*.*

(2) ***Vraisemblance régulière :*** *Pour tout* $\theta \in \Theta$ *et* $x \in \mathcal{A}$*, le gradient* $\nabla_\theta \log p(x;\theta)$ *existe et* $\mathbb{E}_\theta \left| \nabla_\theta \log p(X;\theta) \right| < \infty$*.*

(3) ***Permutabilité*** $\boldsymbol{\nabla_\theta / \int_{\mathcal{X}}}$ ***pour les statistiques intégrables*** *: Si* $S : \mathcal{X} \mapsto \mathbb{R}$ *est une statistique telle que*

$$\mathbb{E}_\theta[|S(X)|] < \infty, \quad \textit{et} \quad \mathbb{E}_\theta[|S(X)\nabla_\theta \log p(X;\theta)|] < \infty, \quad \forall \theta \in \Theta, \tag{3.4}$$

alors, la fonction $\theta \mapsto \mathbb{E}_\theta[S(X)]$ *est différentiable et les opérations de dérivation et d'intégration peuvent être échangées :*

$$\nabla_\theta \int_{\mathcal{X}} S(x) p(x;\theta) \mu(dx) = \int_{\mathcal{X}} S(x) \nabla_\theta p(x;\theta) \mu(dx).$$

La condition (3) est pratique pour établir les résultats qui suivent mais sa vérification rigoureuse est difficile. Pour l'instant, notons simplement que les familles exponentielles, présentées plus bas (section 3.2.4) permettent de mettre en évidence une classe importante de modèles réguliers (voir proposition 3.2.11).

Nous commençons par supposer que $d = 1$ (i.e. $\Theta \subseteq \mathbb{R}$) afin de simplifier la présentation des résultats. L'extension au cas $d > 1$ est faite dans la partie 3.2.3. Lorsque le modèle est régulier, nous pouvons définir la *quantité d'information de Fisher* par

$$I(\theta) \stackrel{\text{def}}{=} \mathbb{E}_\theta \left\{ \left(\frac{\partial \log p}{\partial \theta}(X;\theta) \right)^2 \right\} = \int \left(\frac{\partial \log p}{\partial \theta}(x;\theta) \right)^2 p(x;\theta)\mu(\mathrm{d}x). \tag{3.5}$$

La quantité $I(\theta)$ existe toujours dans le cas d'un modèle régulier, même si elle peut être égale à $+\infty$. La grandeur $I(\theta)$ est une *quantité d'information* au sens de la théorie de l'information. Nous n'élaborerons pas sur ce point, mais nous renvoyons le lecteur intéressé au livre de Cover et Thomas (1991). La famille de variables aléatoires $\frac{\partial \log p}{\partial \theta}(X;\theta)$ $\theta \in \Theta$ s'appelle le *score*.

Une première interprétation (très heuristique) du score et de l'information de Fisher est la suivante : On a mentionné au chapitre introductif que la vraisemblance mesure ... la vraisemblance que θ soit le paramètre de la loi ayant généré l'observation x. Le score (à x fixé) est la dérivée de la log-vraisemblance. Intuitivement, il mesure la possibilité de discriminer entre différents θ au vu d'une observation (dans un contexte où l'on retiendrait le θ dont la vraisemblance est plus élevée, comme dans la section 2.4). À (x, θ_0) fixé, si le score, vu comme une fonction de θ est « plat », on aura du mal à décider si θ_0 est « meilleur » qu'un de ses voisins. Autrement dit, x apporte peu d'information sur θ_0. Ainsi, la dérivée du score est porteuse d'information, et l'information de Fisher est justement l'espérance de cette quantité élevée au carré. Elle représente la quantité d'information moyenne qu'on peut attendre d'une observation (générée selon θ_0). Attention : cette explication n'est pas rigoureuse pour l'instant, la « vraie » raison de l'utilisation de ces quantités est qu'elles apparaissent dans la borne de Cramér-Rao.

Le résultat élémentaire suivant indique que le score est d'espérance nulle sous la *loi du vrai paramètre*, c'est-à-dire, quand X suit lui-même la loi P_θ.

Lemme 3.2.2
Supposons que le modèle est régulier. Alors

$$\mathbb{E}_\theta \left(\frac{\partial \log p}{\partial \theta}(X;\theta) \right) = 0 \,. \tag{3.6}$$

Démonstration. La condition 2 dans la définition 3.2.1 du modèle régulier assure qu'en prenant $T(x) \equiv 1$, les hypothèses d'intégrabilité (3.4) sont satisfaites. Ainsi, d'après la propriété 3 d'un modèle régulier, l'échangeabilité des opérations d'intégration et de dérivation est possible : $\int_{\mathcal{X}} \partial_\theta p(x,\theta)\mu(\mathrm{d}x) = \partial_\theta \int_{\mathcal{X}} p(x,\theta)\mu(\mathrm{d}x)$. Cette dernière quantité est nulle car $\int_{\mathcal{X}} p(x,\theta)\mu(\mathrm{d}x) \equiv 1$. Ainsi,

$$\begin{aligned} \mathbb{E}_\theta \left(\frac{\partial \log p}{\partial \theta}(X;\theta) \right) &= \int \left\{ \frac{\partial p}{\partial \theta}(x;\theta) \big/ p(x;\theta) \right\} p(x;\theta)\mu(dx), \\ &= \int \frac{\partial p}{\partial \theta}(x;\theta)\mu(dx) = \frac{\partial}{\partial \theta} \int p(x;\theta)\mu(dx) = 0. \end{aligned}$$

■

Cette propriété implique en particulier que $I(\theta)$ est la variance du score sous la loi du vrai paramètre :

$$I(\theta) = \mathbb{V}\mathrm{ar}_\theta\left(\frac{\partial \log p}{\partial \theta}(X;\theta)\right). \tag{3.7}$$

Proposition 3.2.3
Soit $(X_1,\dots,X_n)$ n v.a. i.i.d. distribuées suivant un modèle $(\mathrm{P}_\theta^{\otimes n}, \theta \in \Theta \subseteq \mathbb{R})$ vérifiant les conditions du théorème 3.2.4. Notons $p(x;\theta)$ la densité de P_θ et $I_1(\theta) = \mathrm{var}_\theta\left[\left(\frac{\partial \log p}{\partial \theta}(X;\theta)\right)^2\right]$. Alors,

$$I(\theta) = nI_1(\theta)\ .$$

Démonstration. C'est une conséquence directe du lemme 3.2.2. En effet :

$$\begin{aligned} I(\theta) = \mathrm{var}_\theta\left(\frac{\partial \log p}{\partial \theta}(X_1,\dots,X_n;\theta)\right) &= \mathrm{var}_\theta\left(\sum_{i=1}^n \frac{\partial \log p}{\partial \theta}(X_i;\theta)\right) \\ &= \sum_{i=1}^n \mathrm{var}_\theta\left(\frac{\partial \log p}{\partial \theta}(X_i;\theta)\right) = nI_1(\theta). \end{aligned}$$

■

3.2.2 Borne de Cramér-Rao : paramètre scalaire

On se place désormais dans le cadre d'un modèle régulier, au sens de la définition 3.2.1. On considère un paramètre d'intérêt $g(\theta)$. On se donne un estimateur non biaisé de $g(\theta)$, c'est à dire, rappelons-le, une statistique $S : \mathcal{X} \to \mathbb{R}$ telle que $\mathbb{E}_\theta[S(X)] = g(\theta)$, pour tout $\theta \in \Theta$. Le résultat principal de cette section (borne de Cramér-Rao) donne une borne inférieure sur la variance de S, donc sur son risque quadratique (puisque S est non-biaisée).

Pour l'instant, supposons que $g(\theta) \in \mathbb{R}$ (le cas multivarié sera présenté plus bas).

Théorème 3.2.4 (Fréchet-Darmois-Cramér-Rao)
Soient $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$ un modèle régulier, $g(\theta)$ un paramètre d'intérêt et soit $S(X)$ une statistique telle que $\mathbb{E}_\theta S(X) = g(\theta)$ et $\mathbb{V}\mathrm{ar}_\theta[S(X)] < \infty$, pour tout $\theta \in \Theta$. Supposons de plus que $0 < I(\theta) < \infty$. Alors,

$$\mathbb{V}\mathrm{ar}_\theta[S(X)] \geq \frac{g'(\theta)^2}{I(\theta)}. \tag{3.8}$$

Démonstration. Remarquons que, par l'inégalité de Cauchy-Schwarz,

$$\mathbb{E}_\theta\left|S(X)\frac{\partial \log p}{\partial \theta}(X;\theta)\right| \leq \sqrt{\mathbb{E}_\theta[S(X)^2]}\sqrt{I(\theta)} < \infty,$$

et nous pouvons donc, d'après la propriété 3 d'un modèle régulier, dériver $g(\theta) = \mathbb{E}_\theta(S(X))$ sous le signe intégral. Ainsi,

$$g'(\theta) = \int S(x)\frac{\partial p}{\partial \theta}(x;\theta)\mu(dx) = \int S(x)\left(\frac{\partial \log p}{\partial \theta}(x;\theta)\right)p(x;\theta)\mu(dx).$$

Puisque le modèle est régulier, le lemme 3.2.2 (3.6) s'applique et l'espérance du score est nulle. Ainsi, le membre de droite peut être vu comme une covariance,

$$g'(\theta) = \mathrm{cov}_\theta\left(S(X), \frac{\partial \log p}{\partial \theta}(X;\theta)\right)$$

L'inégalité de Cauchy-Schwartz appliqué aux variables aléatoires $S(X)$ et $\frac{\partial \log p}{\partial \theta}(X;\theta)$ nous donne

$$|g'(\theta)|^2 \leq \mathrm{var}_\theta(S(X)) \, \mathrm{var}_\theta\Big(\frac{\partial \log p}{\partial \theta}(X;\theta)\Big),$$

et la preuve est conclue en remarquant $\mathrm{var}_\theta\Big(\frac{\partial \log p}{\partial \theta}(X;\theta)\Big) = I(\theta)$. ■

Corollaire 3.2.5
Supposons que les conditions du théorème 3.2.4 soient satisfaites avec $g(\theta) \equiv \theta$, c'est-à-dire, que S soit un estimateur sans biais et régulier du paramètre θ. Alors,

$$\mathrm{var}_\theta[S(X)] \geq I^{-1}(\theta), \quad \forall \theta \in \Theta.$$

Cette borne est appelée *borne de Cramér–Rao* ou encore *borne de Darmois-Fréchet*.

Exemple 3.3 (Estimation de la moyenne d'un échantillon gaussien)**:**
Soit $(X_1, \ldots, X_n)$ un n-échantillon d'une loi $\mathcal{N}(\theta, \sigma_0^2)$, $\sigma_0 > 0$ connu. Considérons l'estimateur $\bar{X}$ de θ. $\bar{X}$ est un estimateur sans biais et

$$\mathrm{var}_\theta[\bar{X}] = \sigma_0^2/n.$$

On verra dans la section 3.2.4 que les conditions de régularité sont satisfaites dans ce modèle. Nous avons $\frac{\partial \log p}{\partial \theta}(X_i;\theta) = \frac{X_i - \theta}{\sigma_0^2}$ et donc $I(\theta) = n/\sigma_0^2$. Par conséquent $\mathrm{var}_\theta[\bar{X}] = I(\theta)^{-1}$, l'estimateur atteint la borne de Cramér-Rao.

Exemple 3.4 (Estimation du paramètre d'une loi de Bernoulli)**:**
Soit $(X_1, \ldots, X_n)$ un n-échantillon d'une loi de Bernoulli. L'estimateur $\bar{X}$ de θ est sans biais et $\mathrm{var}_\theta[\bar{X}] = \theta(1-\theta)/n$. Pour cet exemple encore, on verra en section 3.2.4 que les hypothèses de régularité du modèle sont satisfaites.

Nous avons $\frac{\partial \log p}{\partial \theta}(x;\theta) = (x-\theta)/\theta(1-\theta)$ et donc $I(\theta) = n/\theta(1-\theta)$. Dans ce modèle encore $\mathrm{var}_\theta[\bar{X}] = 1/I(\theta)$.

Sous des hypothèses de régularité appropriées, nous pouvons aussi écrire l'information de Fisher sous la forme :

$$I(\theta) = -\mathbb{E}_\theta\left(\frac{\partial^2 \log p}{\partial \theta^2}(X;\theta)\right) \tag{3.9}$$

Cette expression est souvent plus simple à calculer.

Proposition 3.2.6
Supposons que $\theta \to p(x;\theta)$ est deux fois différentiable,

$$\int \left|\frac{\partial^2 p}{\partial \theta^2}(x;\theta)\right| \mu(dx) < \infty, \quad \text{et} \quad \frac{\partial^2}{\partial \theta^2}\int p(x;\theta)\mu(dx) = \int \frac{\partial^2 p}{\partial \theta^2}(x;\theta)\mu(dx).$$

Alors, la quantité d'information de Fisher est donnée par (3.9).

Démonstration. Un calcul direct montre que :

$$\frac{\partial^2 \log p}{\partial \theta^2}(x;\theta) = -\left[\frac{\partial p}{\partial \theta}(x;\theta) \big/ p(x;\theta)\right]^2 + \frac{\partial^2 p}{\partial \theta^2}(x;\theta) \big/ p(x;\theta).$$

En remarquant que les hypothèses de la proposition permettent de permuter intégrale et dérivation, on a :

$$\mathbb{E}_\theta\left(\partial_\theta^2 p(X;\theta)/p(X;\theta)\right) = \frac{\partial^2}{\partial\theta^2}\int p(x;\theta)\mu(dx) = 0.$$

D'où le résultat annoncé. ■

Une application simple du théorème 3.2.4 est qu'il permet, dans certains cas de montrer qu'un estimateur est U.V.M.B. (voir définition 3.1.2). On dira qu'un estimateur $\widehat{g}$ de $g(\theta)$ est un estimateur *efficace* de $g(\theta)$ s'il est sans biais, soit $\mathbb{E}_\theta[\widehat{g}(X)] = g(\theta)$, et si sa variance atteint la borne de Cramér-Rao

$$\mathrm{var}_\theta(\widehat{g}(X)) = \frac{g'(\theta)^2}{I(\theta)} \quad \text{pour tout } \theta \in \Theta.$$

Corollaire 3.2.7
Sous les conditions du théorème 3.2.4, un estimateur efficace est nécessairement U.V.M.B.

La réciproque de ce corollaire est fausse. Il est en effet possible de trouver des estimateurs U.V.M.B. qui ne soient pas efficaces.

3.2.3 Borne de Cramér-Rao : paramètre vectoriel

Nous allons maintenant étendre les notions étudiées ci-dessus au cas d'un paramètre multi-dimensionnel, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$. Dans le cas vectoriel, l'information de Fisher est une matrice $d \times d$, définie par

$$I(\boldsymbol{\theta}) = [I_{i,j}(\boldsymbol{\theta})]_{1\le i,j\le d}, \quad I_{i,j}(\boldsymbol{\theta}) = \mathbb{E}_\theta\left(\frac{\partial \log p}{\partial\theta_i}(X;\theta)\frac{\partial \log p}{\partial\theta_j}(X;\theta)\right). \tag{3.10}$$

De façon similaire au cas scalaire, nous avons

Lemme 3.2.8
Supposons que le modèle statistique est régulier. Alors

$$\mathbb{E}_{\boldsymbol{\theta}}\left(\frac{\partial \log p}{\partial\theta_i}(X;\theta)\right) = 0, \quad 1 \le i \le d, \tag{3.11}$$

$$I_{i,j}(\boldsymbol{\theta}) = \mathrm{cov}_{\boldsymbol{\theta}}\left(\frac{\partial \log p}{\partial\theta_i}(X;\theta), \frac{\partial \log p}{\partial\theta_j}(X;\theta)\right), \quad 1 \le i,j \le d. \tag{3.12}$$

La preuve est identique au cas scalaire. On peut réécrire de façon plus compacte les relations précédentes sous la forme

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\nabla_{\boldsymbol{\theta}} \log p(X;\boldsymbol{\theta})\right] = 0, \quad I(\boldsymbol{\theta}) = \mathbb{V}\mathrm{ar}_{\boldsymbol{\theta}}(\nabla_{\boldsymbol{\theta}} \log p(X;\boldsymbol{\theta})),$$

où $\mathrm{Var}(\mathbf{Y})$ est la matrice de variance covariance du vecteur $\mathbf{Y}$.

Proposition 3.2.9 — *Si $(X_1, \ldots, X_n)$ sont des v.a. i.i.d., alors l'information de Fisher associée à $X = (X_1, \ldots, X_n)$ est $I(\boldsymbol{\theta}) = nI_1(\boldsymbol{\theta})$, où $I_1(\boldsymbol{\theta})$ est l'information associée à X_1,*

— *Si $\theta \mapsto p(x,\theta)$ est deux fois différentiable en tout x et si*

$$\int |\nabla^2_{\boldsymbol{\theta}} p(x;\theta)|\mu(dx) < \infty \quad et \quad \nabla^2_{\boldsymbol{\theta}} \int (p)(x;\theta)\mu(dx) = \int \nabla^2_{\theta} p(x;\theta)\mu(dx),$$

alors

$$I(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}}[\nabla^2_{\boldsymbol{\theta}} \log p(X;\theta)], \quad , I_{i,j}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2 \log p}{\partial\theta_i \partial\theta_j}(X;\theta)\right].$$

Exemple 3.5 (Information de Fisher pour une v.a. gaussienne)**:**
Soit X une v.a. de loi $\mathcal{N}(\mu,\sigma^2)$:

$$p(x;\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log\sigma^2\right). \tag{3.13}$$

Les dérivées partielles de $\ell(x;\theta) = \log p(x;\theta)$ par rapport à μ et à σ^2 sont égales à :

$$\nabla_\mu \ell(x;\theta) = \frac{x-\mu}{\sigma^2},$$
$$\nabla_{\sigma^2} \ell(x;\theta) = \frac{(x-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2}.$$

En utilisant le fait que pour v.a. Z de loi $\mathcal{N}(0,1)$, $E[Z^{2n-1}] = 0$ et $E[Z^{2n}] = \prod_{j=1}^n (2j-1)$, il vient

$$I_{11}(\theta) = \frac{1}{\sigma^4}\mathbb{E}_\theta\left[(X_i-\mu)^2\right] = \frac{1}{\sigma^2},$$
$$I_{12}(\theta) = \mathbb{E}_\theta\left[\frac{(X_i-\mu)^3}{2\sigma^6} - \frac{X_i-\mu}{2\sigma^4}\right] = 0,$$
$$I_{22}(\theta) = \frac{1}{4\sigma^4}\mathbb{E}_\theta\left[\left(\frac{X_i-\mu}{\sigma}\right)^4 - 2\left(\frac{X_i-\mu}{\sigma}\right)^2 + 1\right] = \frac{1}{2\sigma^4},$$

d'où l'expression de la matrice d'information de Fisher.

Théorème 3.2.10
On se place dans un modèle régulier. Soit S une statistique à valeurs réelles telle que $\mathbb{V}\mathrm{ar}_{\boldsymbol{\theta}}[S(X)] < \infty$, pour tout $\theta \in \Theta$. Supposons que $0 < I(\boldsymbol{\theta}) < \infty$ et notons $g(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[S]$. Alors $\boldsymbol{\theta} \to g(\boldsymbol{\theta})$ est différentiable et

$$\mathbb{V}\mathrm{ar}_{\boldsymbol{\theta}}[S(X)] \geq \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta})^\top I(\boldsymbol{\theta})^{-1} \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}). \tag{3.14}$$

Comme dans le cas d'un paramètre scalaire, la preuve est une conséquence directe de l'inégalité de Cauchy-Schwarz.

3.2.4 Cas des famille exponentielle

De nombreux modèles sont réguliers au sens de la définition 3.2.1 et admettent des statistiques régulières (sous une condition facile à vérifier d'intégrabilité), en particuliers les modèles de type « famille exponentielle », définis au chapitre 2, section 2.5.

Une des nombreuses propriétés intéressantes de la famille exponentielle (qu'on ne détaillera pas dans ce cours) est la propriété de régularité suivante pour les statistiques intégrables (propriété admise) :

Proposition 3.2.11 (Régularité dans une famille exponentielle)
On considère une famille exponentielle de paramétrisation naturelle $\{\mathrm{P}_\eta, \eta \in \mathcal{E}\}$ *(voir la remarque 2.5.2). Si* $\mathcal{E}$ *est ouvert et si* $S : \mathcal{X} \to \mathbb{R}$ *est une statistique telle que* $\mathbb{E}_\eta(|S(X)|) < \infty$ *pour tout* $\eta \in \mathcal{E}$*, alors l'intégrale* $\mathbb{E}_\eta(S(X))$ *est infiniment dérivable par rapport à chaque composant de* η *et les dérivées partielles peuvent être calculées sous le signe somme.*

Démonstration. idée de la preuve La première étape est de montrer que la constante de normalisation $A(\eta) = -\log \int e^{\langle \eta, T(x)\rangle} h(x)\mu(\mathrm{d}x)$ est infiniment dérivable. Pour cela on montre que si η est à l'intérieur de $\mathcal{E}$, la fonction génératrice des moments de T sous la loi P_η existe pour $t \in \mathbb{R}^p$ suffisamment petit et est donnée par $\mathbb{E}_\eta(e^{\langle t, T(x)\rangle}) = \exp\left(A(\eta + t) - A(\eta)\right)$. L'existence de l'espérance implique (voir Foata and Fuchs [1996], chapitre 13, théorème 13.1) que cette fonction est analytique sur un voisinage de 0. Ainsi, A est infiniment dérivable.

On considère maintenant une composante de η, par exemple η_1. Il faut montrer que la fonction $\eta_1 \mapsto J(\eta_1) = \int S(x) e^{\eta_1 T_1(x) + \sum_{j=2}^p \eta_j T_j(x)} h(x)\mu(\mathrm{d}x)$ est dérivable par rapport à η_1 sous le signe somme, ce qui s'obtient par convergence dominée en considérant la limite de $J(\eta_1+\delta) - J(\eta_1)]/|\delta|$ (pour la domination on utilise la convexité de l'exponentielle). L'argument est le même pour les dérivées d'ordre supérieures, et le caractère $\mathcal{C}^\infty$ s'obtient par récurrence.

La preuve détaillée de ce résultat est donnée dans Lehmann [1959], chapitre 2.

■

En pratique, on pourra utiliser le corollaire suivant pour montrer qu'un modèle exponentiel est régulier et qu'une statistique est régulière dans ce modèle.

Corollaire 3.2.12
Soit $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta \subset \mathbb{R}^d\}$ *une famille exponentielle au sens de la définition 2.5.1. Si* Θ *est ouvert dans* $\mathbb{R}^d$ *et si* $\theta \mapsto \eta(\theta)$ *est continûment différentiable sur* Θ*, alors le modèle est régulier au sens de la définition 3.2.1.*

Ce corollaire s'applique en particulier dans modèles exponentiels cités en exemple à la section 2.5, si l'on exclut les valeurs dégénérées des paramètres, c'est-à-dire en prenant $\Theta =]0, 1[$ pour les modèles de Bernoulli et binomial et $\Theta = \{(\mu, \sigma^2)\} = \mathbb{R} \times \mathbb{R}_+^*$ pour le modèle Gaussien.

Pour conclure ce chapitre, la borne de Cramér-Rao (Théorème 3.2.4) permet de montrer, dans certains cas (lorsque l'estimateur considéré atteint la borne de Cramér-Rao), que des estimateurs sans biais sont U.V.M.B. Cependant, tous les estimateurs U.V.M.B. n'atteignent pas nécessairement la borne en question, c'est-à-dire, tous les estimateurs U.V.M.B. ne sont pas nécessairement efficaces.

Chapitre 4

Optimalité des décisions : cadre classique et cadre bayésien

4.1 Difficultés liées à la minimisation uniforme du risque

Considérons un modèle statistique $\mathcal{P}$ sur l'espace d'observation $\mathcal{X}$, un espace d'actions $\mathcal{A}$, une fonction de perte $L : \mathcal{P} \times \mathcal{A} \to \mathbb{R}_+$ et son risque R associé, défini par (1.5).

Pour définir une hiérarchie entre deux décisions $\delta : \mathcal{X} \to \mathcal{A}$ et $\delta' : \mathcal{X} \to \mathcal{A}$ qui ne dépendent que du modèle, et non de la loi inconnue P_θ, il serait naturel de choisir δ dès lors que

$$R(\theta, \delta) \leq R(\theta, \delta') \quad \text{pour tout} \quad \theta \in \Theta\,. \tag{4.1}$$

Un estimateur δ préférable à tout autre procédure de décision δ' au sens de (4.1) sera appelé *uniformément optimal*. Malheureusement la relation d'ordre ainsi définie sur les procédures de décision est une relation d'ordre partiel, c'est-à-dire qu'elle ne permet pas forcément de comparer toute paire de décisions $\{\delta, \delta'\}$. Une conséquence fâcheuse est qu'il n'existe pas nécessairement de décision uniformément optimale. Par exemple, considérons le modèle $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta = \mathbb{R}\}$, où P_θ est un loi de densité gaussienne de moyenne θ et de variance égale à $1 : X = \theta + \epsilon$, $\epsilon \sim \mathcal{N}(0, 1)$. Considérons le problème de l'estimation de θ. L'espace des actions est $\mathcal{A} = \mathbb{R}$ et on choisit la fonction de perte quadratique. Considérons l'estimateur $\widehat{\theta}(X) = 0$, qui ignore l'observation. Le risque de cet estimateur est $\mathbb{E}_\theta[(0-\theta)^2] = \theta^2$. Cette procédure est la seule [1] qui présente un risque nul à $\theta = 0$ puisque $\mathbb{E}_0[\delta(X)^2] = 0$ implique que $\delta(X) = 0$ p.s. Cet exemple peut paraître troublant car on propose un estimateur qui n'a pas de sens mais est optimal pour un θ particulier ($\theta = 0$). Un exemple plus intéressant est donné ci-après.

Exemple 4.1 (Estimateur de la moyenne à rétrecissement)**:**
On considère un modèle $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$ pour l'observation $\mathbf{X} = (X_1, \ldots, X_n)$ échantillon i.i.d.de loi P_θ. On veut estimer la moyenne $\mu \stackrel{\text{def}}{=} g(\theta) = \mathbb{E}_\theta[X_1]$ sous l'hypothèse $\sigma^2 \stackrel{\text{def}}{=} \mathbb{E}_\theta[X_1^2] < \infty$. On considère l'estimateur à rétrécissement

$$\tilde{X}_n(h) = h\bar{X}_n, \qquad \bar{X}_n = n^{-1}\sum_{i=1}^{n} X_i, \quad \text{moyenne empirique}$$

Le risque quadratique de l'estimateur à rétrécissement est donné par :

$$R(\theta, \tilde{X}_n(h)) \stackrel{\text{def}}{=} \mathbb{E}_\theta\left[(\tilde{X}_n - \mu)^2\right] = \frac{h^2\sigma^2}{n} + \mu^2(1-h)^2.$$

1. au sens presque sûr (p.s.)

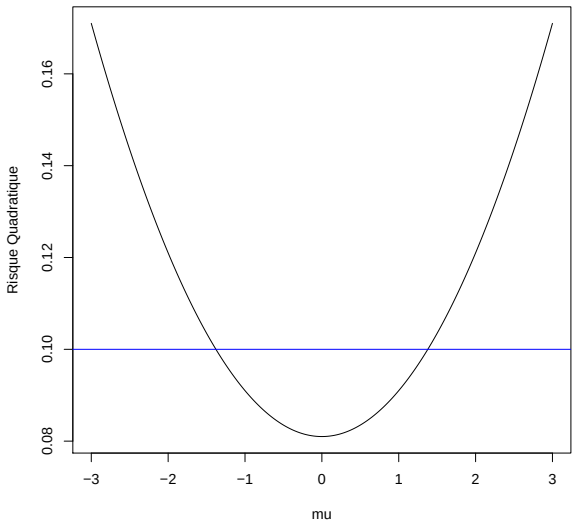


FIGURE 4.1 – Risque quadratique de l'estimateur à rétrécissement (ligne noire) et de la moyenne empirique (ligne bleue) en fonction de μ, pour $h = 0.9, \sigma^2 = 1, n = 10$.

La figure 4.1 montre que le risque de l'estimateur à rétrécissement est plus faible que celui de la moyenne empirique pour des μ proches de 0, mais pas pour les grandes valeurs de μ.

4.2 Optimalité du risque sous contrainte

On a vu qu'on ne pouvait pas systématiquement définir une procédure optimale au sens uniforme donné par (4.1). De façon plus ou moins miraculeuse, dans certains cas, une procédure optimale peut être construite si, dans le critère (4.1), on impose à δ et δ' d'appartenir à des classes particulières.

Exemple 4.2 (Contrainte d'invariance à la translation)**:**
Considérons le modèle d'observation $X_i = U_i + \theta$, où $i \in \{1, \dots, n\}$ et où U_i sont n v.a. i.i.d. centrées. On cherche à estimer $\theta \in \Theta = \mathbb{R}$; l'espace des actions est $\mathcal{A} = \mathbb{R}$ et nous utilisons la perte quadratique. On dit que l'estimateur $\widehat{\theta}(x_1, \dots, x_n)$ est invariant par translation si

$$\widehat{\theta}(x_1 + a, \dots, x_n + a) = a + \widehat{\theta}(x_1, \dots, x_n).$$

Si l'observation $(x_1, \dots, x_n)$ conduit à l'estimateur $\widehat{\theta}(x_1, \dots, x_n)$ alors l'observation translatée de la quantité constante a, conduit à $\widehat{\theta}(x_1, \dots, x_n) + a$.

Cette contrainte d'invariance exclut notamment d'estimer θ par une constante, par exemple $\widehat{\theta} = 0$. On voit aussi que cette contrainte suffit pour résoudre le problème posé par les estimateurs à rétrécissement (exemple 4.1) puisque, dans cet exemple, le seul estimateur $\tilde{X}_n(h)$ obéissant à la contrainte d'invariance à la translation est l'estimateur de la moyenne empirique $(h = 1)$.

Dans l'exemple 4.2 (modèle de translation, à mettre en relation avec l'exemple 1.2 ii), si un estimateur $\widehat{\theta}$ vérifie la contrainte d'invariance à la translation et admet une espérance

finie, alors le biais d'estimation $\mathbb{E}_\theta[\widehat{\theta}] - \theta$ ne dépend pas de θ. Dans ce cas, il est possible et préférable de le fixer à zéro. La contrainte d'invariance devient alors une contrainte d'*absence de biais*. Nous avons vu au chapitre 3 que cette contrainte d'absence de biais permettait dans certains cas d'exhiber des estimateurs optimaux (les estimateurs U.V.M.B), par exemple les estimateurs qui atteignent la borne de Cramér-Rao. Remarquons immédiatement qu'*a contrario*, l'absence de biais ne correspond pas à une contrainte d'invariance à la translation si le paramètre θ n'est pas lui-même un paramètre de translation de la loi.

Il faut bien comprendre que l'utilisation de contraintes est purement *simplificatrice*, et nullement justifiée par une quelconque amélioration de la procédure : l'estimateur à rétrécissement est écarté par la contrainte d'absence de biais alors qu'il est dans certaines situations préférable à l'estimateur de la moyenne empirique. De ce point de vue, les approches précédentes ne donnent en général que des réponses partielles, utiles pour développer une théorie applicable en pratique.

4.3 Risque minimax

Une approche pour définir une relation d'ordre totale entre les décisions sans imposer de contrainte sur les procédures est d'uniformiser le risque en considérant le *pire* risque obtenu quand θ parcourt Θ, on obtient le *risque uniforme* (ou risque maximum) :

$$\sup\left\{R(\theta,\delta):\ \theta\in\Theta\right\}\ , \tag{4.2}$$

quantité appartenant à $\overline{\mathbb{R}}_+$ qui ne dépend, pour un modèle statistique donné, plus que de la procédure de décision δ et qui permet de comparer n'importe quelle paire de procédures entre elles. On obtient ainsi une hiérarchie des décisions : nous choisissons δ plutôt que δ' si

$$\sup\left\{R(\theta,\delta):\ \theta\in\Theta\right\} \leq \sup\left\{R(\theta,\delta'):\ \theta\in\Theta\right\}\ . \tag{4.3}$$

Le risque minimax (*mini*mum du risque *max*imum) est alors défini par

$$R_{\text{minimax}} = \inf_\delta \sup_\theta R(\theta,\delta)\ ,$$

où l'infinimum est pris sur l'ensemble des procédures de décision $\delta : \mathcal{X} \to \mathcal{A}$ et le supremum sur l'espace des paramètres $\theta \in \Theta$ du modèle $\mathcal{P}$. En pratique le calcul du risque minimax et la recherche de procédures de décision approchant ce risque sont très difficiles à déterminer.

Il y a des alternatives à l'approche minimax pour comparer les risques des procédures de décision. Si l'espace des paramètres Θ est tel que l'on peut définir une mesure π sur Θ, on peut remplacer le risque uniforme (4.2) par un risque intégré, ce qui permet de "moyenner" le risque sur tous les θ possibles. Pour cela on considère le risque intégré

$$\int_\Theta R(\theta,\delta)\ \pi(\mathrm{d}\theta)\ ,$$

pour lequel il est parfois plus facile de déterminer la procédure δ qui le minimise. Ce sera l'approche utilisée dans la modélisation bayésienne, que nous introduisons plus précisément ci-dessous.

4.4 La modélisation bayésienne

Introduction

Nous avons jusqu'ici supposé que les observations X pouvaient nous renseigner sur leur loi P_θ, en faisant l'hypothèse préliminaire que P_θ appartient à une famille $\mathcal{P}$ donnée (le modèle). La définition de cette famille, ce qu'on a appelé le *modèle statistique* constitue dans ce cas la connaissance *a priori* des propriétés statistiques des données. Autrement dit, la donnée d'un modèle fixe la connaissance a priori sous la forme d'une famille de probabilités $(\mathrm{P}_\theta, \theta \in \Theta)$ *possibles* fixée. Il existe des situations pour lesquelles on peut affiner cette connaissance a priori en décrivant quels paramètres θ sont les plus *probables*, c'est-à-dire en définissant une mesure de probabilité sur l'espace des paramètres. Cette mesure que l'on fixe *avant* d'observer les données représente le degré de crédibilité accordé par le statisticien à telle ou telle valeur de θ (ou région de Θ dans le cas non dénombrable) avant d'avoir réalisé l'expérience, c'est-à-dire la connaissance *a priori* du statisticien concernant le problème statistique envisagé. Il peut être relativement uniforme en l'absence d'information, ou au contraire concentré sur de petites régions de Θ si le contexte (données historiques pré-existantes, connaissance d'expert ...) le permet. Dans l'exemple 1.1, supposons que nous disposions d'un historique du nombre d'objets défectueux dans les échantillons de test. Cet historique nous permet d'obtenir une information a priori (c'est-à-dire, avant d'examiner l'échantillon courant) sur la fréquence $\{\pi_0, \ldots, \pi_N\}$ du nombre d'objets défectueux dans la population. Dans une telle situation, il est raisonnable de se donner comme mesure de crédibilité *a priori* la distribution donnée par les fréquences relatives, π_i. On a ainsi défini une loi de probabilité π sur l'espace des paramètres, $\pi(i/n) = \pi_i$. On peut donc voir le « vrai » θ (celui ayant servi à générer les données) comme une réalisation d'une variable aléatoire $\boldsymbol{\theta}$ de loi π. On vient de donner un exemple particulier de *modèle bayésien* : la loi jointe du couple $(\boldsymbol{\theta}, X)$ (paramètre et observations) est donnée par :

$$\mathbb{P}\left(\boldsymbol{\theta} = i/N, X = k\right) = \mathbb{P}\left(X = k | \boldsymbol{\theta} = i/N\right) \mathbb{P}\left(\boldsymbol{\theta} = i/N\right), \tag{4.4}$$

$$= \pi_i \frac{\binom{i}{k}\binom{N-i}{n-k}}{\binom{N}{n}}, \quad k \geq i, n-k \geq N-i. \tag{4.5}$$

L'objectif de la modélisation bayésienne est d'utiliser l'observation X pour mettre à jour la connaissance sur $\boldsymbol{\theta}$. Ceci s'effectue en *conditionnant* la loi de θ à l'observation $X = x$, c'est-à-dire, en calculant la loi conditionnelle de $\boldsymbol{\theta}$ sachant l'observation X. Dans notre exemple, la loi conditionnelle de $\boldsymbol{\theta}$ sachant $X = k$ est donnée par

$$\begin{aligned}\mathbb{P}\left(\boldsymbol{\theta} = i/N | X = k\right) &\overset{\text{def}}{=} \frac{\mathbb{P}(\theta = i/n, X = k)}{\mathbb{P}(X = k)} \\ &= \frac{\mathbb{P}\left(X = k | \boldsymbol{\theta} = i/N\right) \mathbb{P}\left(\boldsymbol{\theta} = i/N\right)}{\sum_{j=0}^{N} \mathbb{P}\left(X = k | \boldsymbol{\theta} = j/N\right) \mathbb{P}\left(\boldsymbol{\theta} = j/N\right)}.\end{aligned}$$

Cette loi conditionnelle sera appelée *loi a posteriori*. Les paragraphes suivants formalisent ces notions dans un cas plus général.

4.4.1 Modèle bayésien

Supposons que nous disposions d'un modèle statistique $\{\mathrm{P}_\theta, \theta \in \Theta\}$. Pour obtenir un modèle bayésien, nous introduisons une variable aléatoire $\boldsymbol{\theta}$, définie sur l'espace $(\Omega, \mathcal{B}(\Omega))$ à

valeurs dans $(\Theta, \mathcal{B}(\Theta))$, où $\mathcal{B}(\Theta)$ est la *tribu des paramètres*. Lorsque l'espace des paramètres est discret $\Theta = \{\theta_1, \theta_2, \ldots\}$, on prend pour $\mathcal{B}(\Theta)$ les parties de Θ. Lorsque $\Theta = \mathbb{R}^d$, on prend pour $\mathcal{B}(\Theta)$ la tribu borélienne. Dans le cas non-paramétrique, il est encore possible de définir une tribu Borélienne dès lors qu'on se donne une définition des ouverts. Dans ce cours, nous nous restreignons au cas paramétrique.

Notons π la loi de $\boldsymbol{\theta}$: π représente l'information dont on dispose sur le paramètre avant que l'expérience ne fournisse les observations. On appelle π la *loi a priori*.

Définition 4.4.1 (Modèle bayésien)**.** *Un modèle bayésien est la donnée de*

(1) *Un modèle statistique* $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$,

(2) *Une tribu des paramètres* $\mathcal{B}(\Theta)$ *et une loi* a priori π *sur* $(\Theta, \mathcal{B}(\Theta))$.

4.4.2 Loi jointe, loi marginale des observations

La donnée d'un modèle bayésien comme dans la définition 4.4.1 permet de définir :

(a) La **loi jointe du couple** $(\boldsymbol{\theta}, X)$ sur $\Theta \times \mathcal{X}$, que l'on notera P_π, donnée par

$$\mathrm{P}_\pi(A \times B) = \mathbb{P}(\boldsymbol{\theta} \in A, X \in B) = \int_A \mathrm{P}_\theta(B)\pi(\mathrm{d}\theta), \qquad A \subset \Theta, B \subset \mathcal{X}. \tag{4.6}$$

ou encore, pour toute fonction $\varphi : \Theta \times \mathcal{X} \to \mathbb{R}^+$ mesurable,

$$\mathbb{E}[\varphi(\boldsymbol{\theta}, X)] = \int_\Theta \left[\int_\mathcal{X} \varphi(\theta, x) \mathrm{P}_\theta(\mathrm{d}x) \right] \pi(\mathrm{d}\theta).$$

(b) La **loi marginale de** X, aussi appelée **marginale a priori**, que nous noterons m^X. C'est la loi de X « en moyenne », après intégration sous la loi *a priori* π,

$$m^X(A) = \mathrm{P}_\pi(\Theta \times A) = \mathbb{P}(X \in A, \boldsymbol{\theta} \in \Theta) = \int_\Theta \mathrm{P}_\theta(A)\pi(\mathrm{d}\theta). \tag{4.7}$$

Lorsque le modèle est dominé par une mesure de référence ν, en notant p_θ la densité de P_θ, *i.e.* $\mathrm{P}_\theta(\mathrm{d}x) = p_\theta(x)\nu(\mathrm{d}x)$, l'équation (4.7) se ré-écrit

$$\begin{aligned} m^X(A) &= \int_\Theta \int_A p_\theta(x)\nu(\mathrm{d}x)\pi(\mathrm{d}\theta). \\ &= \int_A \underbrace{\left(\int_\Theta p_\theta(x)\pi(\mathrm{d}\theta) \right)}_{m(x)} \nu(\mathrm{d}x) \qquad \text{(Fubini)} \end{aligned}$$

Ainsi, la loi marginale de X admet une densité par rapport à ν, donnée par

$$m(x) = \int_\Theta p_\theta(x)\pi(\mathrm{d}\theta). \tag{4.8}$$

4.4.3 Conditionnement

Au vu de (4.6), P_θ est la *loi conditionnelle de* X *sachant* $\{\boldsymbol{\theta} = \theta\}$. De manière intuitive, ceci signifie que P_θ décrit le comportement probabiliste de X à $\boldsymbol{\theta} = \theta$ fixé. Nous donnons ci-dessous une définition précise d'une loi conditionnelle.

Définition 4.4.2 (Loi conditionnelle)**.** *Soit un couple de variables aléatoires* (X, Y) *défini de* $(\Omega, \mathcal{F}, \mathbb{P})$ *dans* $\mathbb{R}^d \times \mathcal{Y}$*. On note* P^Y *la loi marginale de* Y*. On appelle* noyau de loi conditionnelle de X sachant Y *toute famille de lois de probabilités* $(\mathrm{P}_{X|y})_{y\in\mathcal{Y}}$ *sur* $\mathbb{R}^d$ *telle que pour tous* $A \subset \mathbb{R}^d$ *et* $B \subset \mathcal{Y}$ *mesurables, l'application* $y \mapsto \mathrm{P}_{X|y}(A)$ *est mesurable et*

$$\mathbb{P}(X \in A, Y \in B) = \int_B \mathrm{P}_{X|y}(A)\mathrm{P}^Y(\mathrm{d}y), \qquad A \subset \mathcal{Y}, B \subset \mathbb{R}^d. \tag{4.9}$$

À y *fixé, on appelle* loi conditionnelle de X sachant $Y = y$ *la mesure de probabilité* $\mathrm{P}_{X|y}(\,\cdot\,)$.

L'équation (4.9) n'est autre que la formule (4.6) définissant la loi jointe dans un modèle bayésien, en prenant $Y = \boldsymbol{\theta}$, $\mathrm{P}^Y = \pi$ et $\mathrm{P}_{X|y} = \mathrm{P}_\theta$. Ainsi, dans un cadre bayésien, la loi P_θ est la loi conditionnelle de X sachant $\{\boldsymbol{\theta} = \theta\}$. Remarquons que, au vu de (4.9), la loi jointe d'un couple est entièrement déterminée par la donnée de la loi marginale de Y et du noyau de loi conditionnelle de X sachant Y. Cette interprétation en termes de lois conditionnelles d'un modèle bayésien justifie la notation suivante :

$$\begin{aligned} \boldsymbol{\theta} &\sim \pi \\ X|\theta &\sim \mathrm{P}_\theta, \end{aligned}$$

qui désigne un modèle bayésien défini par le modèle $\{\mathrm{P}_\theta, \theta \in \Theta\}$, muni de la loi *a priori* π sur Θ.

L'intérêt de ce formalisme n'est bien sûr pas de remarquer que P_θ est une loi conditionnelle, mais d'inverser le sens du conditionnement pour calculer la loi de $\boldsymbol{\theta}$ sachant $X = x$, notée $\mathrm{P}_{\boldsymbol{\theta}|x}$ ou plus simplement $\pi(\,\cdot\,|x)$, cette dernière représentant la connaissance sur $\boldsymbol{\theta}$ dont on dispose après avoir observé X. Avant tout, on a besoin de s'assurer de la possibilité d'inverser ce conditionnement et d'avoir les outils pour calculer $\mathrm{P}_{\boldsymbol{\theta}|x}$. On admet pour cela les deux résultats suivants :

Proposition 4.4.3 (Loi conditionnelle : existence et unicité)
Soit (X, Y) *un couple aléatoire comme dans la définition 4.4.2.*

(1) *Il existe un noyau de loi conditionnelle de* X *sachant* Y*, c'est-à-dire une famille de lois de probabilités* $(\mathrm{P}_{X|y})_{y\in\mathcal{Y}}$*, qui vérifie (4.9).*

(2) *Cette famille est définie de manière unique en dehors d'un ensemble* $N \subset \mathcal{Y}$ *de* P^Y*-mesure nulle (*i.e. $\mathrm{P}^Y(N) = 0$*). On peut donc parler* du *noyau de loi conditionnelle de* X *sachant* Y*, et (presque sûrement) de* la *loi conditionnelle de* X *sachant* $Y = y$*.*

Proposition 4.4.4 (Loi conditionnelle : Caractérisation)
Soit (X, Y) *comme dans la définition 4.4.2. Une famille de lois de probabilités* $(\mathrm{P}_{X|y})_{y\in\mathcal{Y}}$ *sur* $\mathbb{R}^d$ *est le noyau de loi conditionnelle de* X *sachant* Y *si et seulement si*

(1) *L'application* $y \mapsto \mathrm{P}_{X|y}(A)$ *est mesurable pour tout* $A \subset \mathbb{R}^d$ *, et*

(2) *Pour toute fonction mesurable* $\varphi : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}^+$*,*

$$\mathbb{E}\left[\varphi(X, Y)\right] = \int_{\mathcal{Y}} \left[\int_{\mathbb{R}^d} \varphi(x, y)\mathrm{P}_{X|y}(\mathrm{d}x)\right] \mathrm{P}^Y(\mathrm{d}y). \tag{4.10}$$

Une construction de la loi conditionnelle, basée sur l'*espérance conditionnelle*, est donnée en appendice (définition A.11.11) et n'est pas nécessaire à la compréhension de ce cours. Dans le cas continu, on peut calculer explicitement la densité de la loi conditionnelle. Cette dernière est appelée *densité conditionnelle*.

Proposition 4.4.5 (Cas continu : expression de la densité conditionnelle)
. *Soit (X, Y) un couple aléatoire comme dans la définition 4.4.2. Supposons que la loi jointe $\mathbb{P}_{X,Y}$ admette une densité $f(x, y)$ par rapport à une mesure produit $\nu \otimes \mu$ sur $\mathbb{R}^d \times \mathcal{Y}$. Notons m^Y la densité de la loi marginale de Y par rapport à μ, $m^Y(y) = \int_{\mathbb{R}^d} f(x, y)\nu(\mathrm{d}x)$. Alors, la loi conditionnelle $\mathrm{P}_{X|y}$ admet une densité par rapport à ν, que l'on notera $p(x|y)$ ou $p_y(x)$, donnée par*

$$p(x|y) = \begin{cases} \dfrac{f(x,y)}{m^Y(y)} & \text{si } m^Y(y) \neq 0, \\ p_0 & \text{si } m^Y(y) = 0, \end{cases} \tag{4.11}$$

ou p_0 est une densité de probabilité arbitraire sur $\mathbb{R}^d$.

Démonstration. Tout d'abord, pour tout y tel que $m^Y(y) \neq 0$, $p(x|y)$ définie comme dans (4.11) est bien une densité de probabilité sur $\mathbb{R}^d$ car elle est positive, mesurable en x et vérifie $\int_{\mathbb{R}^d} p(x|y)\nu(\mathrm{d}x) = 1$. On utilise ensuite la caractérisation de la proposition 4.4.4. On va montrer que la famille de lois de probabilités $\mathrm{P}_{X|y}$ définie par $\mathrm{P}_{X|y}(\mathrm{d}x) = p(x|y)\nu(\mathrm{d}x)$ vérifie (4.10).

Soit $\varphi : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}^+$, une fonction mesurable. Alors

$$\begin{aligned}
\mathbb{E}[\varphi(X,Y)] &= \int_{\mathbb{R}^d \times \mathcal{Y}} \varphi(x,y) f(x,y)\ \nu \otimes \mu(\mathrm{d}x, \mathrm{d}y) \\
&= \int_{\mathcal{Y}} \left[\int_{\mathbb{R}^d} \varphi(x,y) f(x,y)\ \nu(\mathrm{d}x) \right] \mu(\mathrm{d}y) \qquad \text{(Fubini)} \\
&= \int_{\{y : m^Y(y) \neq 0\}} \left[\int_{\mathbb{R}^d} \varphi(x,y) p(x|y) \nu(\mathrm{d}x) \right] m^Y(y) \mu(\mathrm{d}y) \\
&+ \int_{\{y : m^Y(y) = 0\}} \left[\int_{\mathbb{R}^d} \varphi(x,y) f(x,y) \nu(\mathrm{d}x) \right] \mu(\mathrm{d}y).
\end{aligned}$$

Le deuxième terme du membre de droite est nul car $m^Y(y) = 0 \Rightarrow f(x, y) = 0$ pour ν-presque tout x. De plus on peut étendre l'intégrale du premier terme à $\mathcal{Y}$ tout entier car l'intégrande est nulle sur l'ensemble $\{y : m^Y(y) = 0\}$. On a donc bien

$$\mathbb{E}[\varphi(X,Y)] = \int_{\mathcal{Y}} \left[\int_{\mathbb{R}^d} \varphi(x,y) p(x|y) \nu(\mathrm{d}x) \right] m^Y(y) \mu(\mathrm{d}y) = \int_{\mathcal{Y}} \left[\int_{\mathbb{R}^d} \varphi(x,y) \mathrm{P}_{X|y}(\mathrm{d}x) \right] \mathrm{P}^Y(\mathrm{d}y).$$

Il reste à voir que pour tout $A \subset \mathbb{R}^d$, la fonction $y \mapsto \mathrm{P}_{X|y}(A)$ est mesurable en tant que fonction de y, ce qui est une conséquence directe du théorème de Fubini, appliqué à f, qui est une fonction mesurable (par rapport à la tribu produit). ■

Remarque 4.4.6. *La formule de la loi conditionnelle (4.11) est à rapprocher de la formule de conditionnement déjà connue dans le cas discret*

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

4.4.4 Loi a posteriori

La notion probabiliste de conditionnement permet de définir rigoureusement la *loi a posteriori* du paramètre $\boldsymbol{\theta}$ dans un modèle bayésien, sachant l'observation X. Dans toute la suite,

pour éviter d'avoir à faire des hypothèses techniques de régularité sur Θ (assurant l'existence d'une telle loi conditionnelle), on suppose $\Theta \subset \mathbb{R}^d$. On a déjà dit informellement que la loi a posteriori devait représenter la connaissance sur θ, après mise à jour de la connaissance a priori π par la donnée $X = x$. Voici une définition mathématique :

Définition 4.4.7. *Soit un modèle bayésien $(\mathcal{P}, \pi)$ donné par un modèle paramétrique $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta \subset \mathbb{R}^p\}$ sur un espace d'observations $\mathcal{X}$ et par un prior π sur Θ.*

La loi a posteriori *est le noyau de loi conditionnelle de $\boldsymbol{\theta}$ sachant X. C'est donc une famille de lois de probabilité indexée par $x \in \mathcal{X}$. On la notera $(\pi(\,\cdot\,|x))_{x\in\mathcal{X}}$.*

Remarque 4.4.8. *La proposition 4.4.3 assure l'existence de la loi a posteriori dans le cadre paramétrique.*

Dans le cas d'un modèle dominé (existence de densités), on peut déterminer la loi a posteriori en écrivant explicitement sa densité. Supposons que $\{\mathrm{P}_\theta,\, \theta \in \Theta\}$ soit un modèle dominé, $\mathrm{P}_\theta(dx) = p_\theta(x)\nu(dx)$ et soit μ une mesure dominant la loi π, et continuons de noter π la densité, $\pi(d\theta) = \pi(\theta)\mu(d\theta)$. La densité jointe du vecteur aléatoire $(\boldsymbol{\theta}, X)$ par rapport à la mesure produit $\mu \otimes \nu$ est alors [2] donnée par :

$$f(\theta, x) = \pi(\theta)p_\theta(x).$$

On déduit directement de la proposition 4.4.5 et de l'expression (4.8) de la densité marginale de X la proposition suivante .

Proposition 4.4.9 (densité a posteriori)
Sous les hypothèses précédentes (modèle dominé), la densité de la loi a posteriori par rapport à la mesure de référence μ est donnée par

$$\pi(\theta|x) = \frac{f(\theta, x)}{m(x)} = \frac{p_\theta(x)\pi(\theta)}{\int_\Theta p_t(x)\pi(t)\mu(\mathrm{d}t)} \tag{4.12}$$

4.4.5 Espérance a posteriori

Supposons que le commanditaire d'une étude demande de fournir une estimation $\widehat{\theta}$ du paramètre θ_0 (le « vrai » paramètre) dont proviennent les données. Comme d'habitude, on suppose que $X \sim \mathcal{P}_{\theta_0}$ avec $\mathrm{P}_{\theta_0} \in \{\mathcal{P}_\theta, \theta \in \Theta \subset \mathbb{R}^d\}$, un modèle statistique paramétrique donné. Supposons également que l'expertise technique du commanditaire permette au statisticien de définir un prior π sur Θ.

Le statisticien se trouve alors dans le cadre de l'estimation ponctuelle : on ne lui demande pas de fournir la loi a posteriori (le commanditaire n'en a peut-être jamais entendu parler) mais un nombre. Une idée naturelle, en supposant que la variance de la loi a posteriori est faible, consiste à fournir comme estimateur l'espérance de $\widehat{\theta}$, sous la loi a posteriori, c'est à dire, si x est la donnée observée,

$$\widehat{\theta}(x) = \int_\Theta \theta\pi(\theta|x)\mu(\mathrm{d}\theta).$$

2. à condition que l'application $\theta \mapsto p_\theta(x)$ soit mesurable, ce qui sera toujours le cas dans les modèles utilisés en pratique.

Plus généralement, si l'on cherche à estimer une grandeur $g(\theta_0) \in \mathbb{R}^p$, il parait raisonnable de prendre comme estimateur l'espérance de $g(\boldsymbol{\theta})$ sous la loi a posteriori,

$$\widehat{g}\,(x) = \int_\Theta g(\theta)\pi(\theta|x)\mu(\mathrm{d}\theta).$$

Remarquons que ces deux estimateurs sont des fonctions des données observées, ce sont donc bien des *statistiques*. On les a construits à partir d'intégrales sous la loi a posteriori. C'est un exemple d'utilisation de la notion d'espérance conditionnelle, définie de manière plus générale comme suit.

Définition 4.4.10 (Espérance conditionnelle)**.** *Soit (X, Y) un couple de variables aléatoires comme dans le définition 4.4.2, à valeurs dans $\mathbb{R}^d \times \mathcal{Y}$, avec $X = (X_1, \ldots, X_d)$, tel que $\mathbb{E}(\sum_{i=1}^d (|X|_i)) < \infty$. Soit $(\mathrm{P}_{X|y})_{y\in\mathcal{Y}}$ la loi conditionnelle de X sachant Y.*

(1) *L'espérance conditionnelle de X sachant $\{Y = y\}$, notée $\mathbb{E}[X|Y = y]$ est la quantité définie(P^Y-presque partout) par l'espérance de X sous la loi conditionnelle $\mathrm{P}_{X|y}$,*

$$\mathbb{E}[X|Y = y] = \int_{\mathbb{R}^d} x\mathrm{P}_{X|y}(\mathrm{d}x).$$

(2) *Soit ϕ la fonction définie presque partout par $\phi(y) = \mathbb{E}(X|Y = y)$.* L'espérance de X sachant Y, *notée $\mathbb{E}(X|Y)$ est la variable aléatoire définie par*

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

Remarque 4.4.11 (Lien avec l'espérance)**.** *L'hypothèse $\mathbb{E}(\sum_{i=1}^d |X|_i) < \infty$ assure que $\mathbb{E}[X|Y = y]$ existe et est finie presque partout, et est intégrable en tant que fonction de y.*

En effet, dans le cas $d = 1$, et si X est une v.a. positive, alors $\mathbb{E}(|X|) < \infty$ si et seulement si (d'après (4.10))

$$\int_{\mathcal{Y}} \underbrace{\int_{\mathbb{R}} x\mathrm{P}_{X|y}(\mathrm{d}x)}_{\mathbb{E}(X|Y=y)} \mathrm{P}^Y(\mathrm{d}y) < \infty,$$

ce qui implique que $\mathbb{E}(X|Y = y)$ est finie presque partout et intégrable sous la loi marginale P^Y de Y. Une deuxième conséquence immédiate et très pratique de (4.10) est la règle de calcul de l'espérance (pour X une v.a. intégrable sous la loi marginale P^X)

$$\mathbb{E}(X) = \mathbb{E}\left[\mathbb{E}\left[X|Y\right]\right]. \tag{4.13}$$

L'extension au cas intégrable s'effectue comme d'habitude en considérant la partie positive et négative, puis le cas multivarié ($d > 1$) se traite en considérant les composantes une à une.

Une construction plus directe, mais plus abstraite de l'espérance conditionnelle, qui n'est pas nécessaire à la compréhension de ce cours, est donnée en appendice A.11.

Dans le cadre de la modélisation bayésienne, la notion d'espérance conditionnelle permet de définir l'espérance a posteriori d'une quantité d'intérêt $g(\theta) \in \mathbb{R}^d$.

Définition 4.4.12 (Espérance a posteriori)**.** *On se donne un modèle bayésien $(\mathcal{P}, \pi)$ où $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta \subset \mathbb{R}^d\}$ est un modèle paramétrique. Soit $g : \theta \mapsto g(\theta) \in \mathbb{R}^p$ une quantité*

*d'intérêt. L'*espérance a posteriori *de* $g(\boldsymbol{\theta})$*, sachant l'observation* $X = x$*, est l'espérance conditionnelle de* $g(\boldsymbol{\theta})$ *sachant* $X = x$*,*

$$\mathbb{E}(g(\boldsymbol{\theta})|X = x) = \int_\Theta g(\theta)\pi_{\theta|x}(\mathrm{d}\theta).$$

Ainsi, dans le cas d'un modèle dominé, soit $\pi(\mathrm{d}\theta) = \pi(\theta)\mu(\mathrm{d}\theta)$ *et* $\mathrm{P}_\theta(\mathrm{d}x) = p_\theta(x)\nu(\mathrm{d}x)$*, l'espérance a posteriori de* $g(\boldsymbol{\theta})$ *est donnée par*

$$\mathbb{E}(g(\boldsymbol{\theta})|X = x) = \int_\Theta g(\theta)\pi(\theta|x)\mu(\mathrm{d}\theta).$$

Résumé

Résumons les idées principales de la modélisation bayésienne :

- Le modèle statistique $\mathcal{P} = \{\mathrm{P}_\theta : \theta \in \Theta\}$ nous donne la loi de X sachant $\{\boldsymbol{\theta} = \theta\}$, et le statisticien définit avant l'expérience une distribution *a priori* π sur Θ. La donnée du couple $(\mathcal{P}, \pi)$ définit un modèle bayésien.
- Après l'expérience consistant à observer X, le résultat de l'analyse est la loi conditionnelle de $\boldsymbol{\theta}$ (le paramètre inconnu qui nous intéresse) sachant les observations $\{X = x\}$, notée $\pi(\,\cdot\,|x)$, et appelée *loi a posteriori*.
- Dans le cas d'un modèle dominé, et si π admet une densité (également notée π), alors la loi a posteriori admet une densité, que l'on note encore $\pi(\theta|x)$, donnée par la formule (4.12). Cette expression de la loi a posteriori par sa densité est appelée *formule de Bayes*.
- La loi a posteriori permet de définir l'espérance a posteriori d'une quantité d'intérêt $g(\theta)$, qu'on utilisera plus tard dans des problèmes d'estimation et de tests statistiques.

Avant de conclure ce paragraphe sur un exemple qui nous permettra d'introduire la notion de familles conjuguées, notons ici que la densité a posteriori appliquée aux observations, c'est-à-dire la fonction $\theta \to \pi(\theta|x)$ dans le cadre bayésien jouera un rôle équivalent à la vraisemblance dans le cadre non-bayésien des modèles dominés.

Exemple 4.3 (Loi de Bernoulli)**:**
Soient $X = (X_1, \ldots, X_n)$ un vecteur de n v.a. i.i.d. de loi Bernoulli P_θ, $\theta \in \Theta = [0, 1]$. Notons π la densité de la loi a priori pour θ par rapport à la mesure de Lebesgue sur $[0, 1]$. La loi a posteriori est donnée par

$$\pi(\theta|x) = \frac{\pi(\theta)\theta^{S_n(x)}(1-\theta)^{n-S_n(x)}}{\int_0^1 \pi(t)t^{S_n(x)}(1-t)^{n-S_n(x)}dt} \tag{4.14}$$

où $S_n(x) := \sum_{i=1}^n x_i$ pour $x = (x_1, \ldots, x_n)$. Remarquons que la loi a posteriori ne dépend des observations qu'à travers la statistique S_n, le nombre total de succès pendant l'expérience.

Pour loi a priori pour θ, nous avons besoin d'une loi dont le support soit inclus dans l'intervalle $[0, 1]$. Parmi les choix possibles de telles lois, il est intéressant de considérer la famille des lois Bêta. Les lois $Beta$ dépendent de deux paramètres α, β et la densité de la loi $\mathcal{B}eta(\alpha, \beta)$ est donnée par :

$$b_{\alpha,\beta}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}, \quad 0 < x < 1,$$

où $B(\alpha,\beta)=\Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ est la fonction Bêta et Γ la fonction Gamma, voir (A.13). En substituant cette expression de la densité dans (4.14), nous obtenons

$$\pi(\theta|y)=\frac{\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}}{B(y+\alpha,n-y+\beta)},$$

et donc la loi a posteriori est encore une loi $\mathcal{B}eta(y+\alpha,n-y+\beta)$. Il s'agit ici d'un phénomène particulier (mais non exceptionnel) : la loi a priori et a posteriori appartiennent à la même famille de loi de probabilité : seuls les paramètres de ces lois sont différentes, traduisant ainsi l'information apportée par l'expérience statistique. De telles lois a priori sont appelées *conjuguées* : la famille des lois Bêta est conjuguée de la famille de loi de Bernoulli. L'utilisation de lois conjuguées simplifie l'inférence bayésienne : bien entendu, si l'on dispose d'informations qui ne sont pas "compatibles" avec les lois a priori conjugués, il est nécessaire d'utiliser d'autres familles de loi a priori. En utilisant les résultats classiques sur les lois Bêta, on montre aisément que la moyenne de $\boldsymbol{\theta}$ sous la loi a posteriori (i.e. l'espérance conditionnelle de $\boldsymbol{\theta}$ sachant $S_n=y$) est donnée par

$$\mathbb{E}\left[\boldsymbol{\theta}|S_n=y\right]=\frac{\alpha+y}{\alpha+\beta+n}.$$

On remarque que cette quantité est dans le segment délimité par la fréquence empirique y/n et la moyenne a priori, $\alpha/(\alpha+\beta)$. La variance a posteriori est donnée par

$$\mathrm{var}[\boldsymbol{\theta}|S_n=y]=\frac{\mathbb{E}[\boldsymbol{\theta}|S_n=y](1-\mathbb{E}[\boldsymbol{\theta}|S_n=y])}{\alpha+\beta+n+1}.$$

Pour des valeurs données de α et β, et lorsque y et n sont grands, on remarque que $\mathbb{E}_\theta[\boldsymbol{\theta}|S_n=y]\simeq y/n$ et $\mathrm{var}[\boldsymbol{\theta}|S_n=y]\simeq n^{-1}(y/n)(1-(y/n))$, qui tend vers 0 à la vitesse $1/n$. Clairement, lorsque la taille de l'échantillon $n\to\infty$, l'influence des paramètres de la loi a priori disparaît.

La loi $\mathcal{B}eta(1,1)$ est la loi uniforme sur $[0,1]$. C'est la loi a priori utilisée par Bayes (1763) et redécouverte indépendamment par Laplace (1800), fondateurs de l'estimation bayésienne, pour l'analyse bayésienne du modèle de Bernoulli. La motivation première de Laplace était de déterminer si le nombre de garçons et de filles à la naissance suivait une loi de Bernoulli de paramètre 0.5. Un total de 241945 filles et de 251527 garçons sont nés à Paris de 1745 à 1770. En appelant "succès" la naissance d'un enfant de sexe féminin, Laplace a montré que

$$\text{pour } n=241945+251527,\ \pi_{\boldsymbol{\theta}|S_n}(\theta\geq 0.5|S_n=241945)=1.15\times 10^{-42},$$

montrant qu'avec une probabilité très proche de 1, $\theta<0.5$. Comme nous l'avons noté ci-dessus, pour des valeurs aussi grandes, l'influence de la loi a priori est tout à fait négligeable, la loi a posteriori étant extrêmement concentrée autour de la valeur $\widehat{\theta}_n=241945/(241945+251527)=0.49$.

La définition générale des familles conjuguées introduite par l'exemple précédent est donnée ci-après.

Définition 4.4.13. *Soient $\mathcal{P}=\{p_\theta:\theta\in\Theta\}$ une famille de densités définies sur $(\mathcal{X},\mathcal{B}(\mathcal{X}))$ et Π une famille de lois définies sur $(\Theta,\mathcal{B}(\Theta))$. Nous dirons que la famille Π est* conjuguée *à la famille $\mathcal{P}$ si, pour tout $\pi\in\Pi$, la loi a posteriori associée au modèle bayésien $p(x|\theta)=p_\theta(x)$ de loi a priori π appartient aussi à Π.*

Cette propriété est particulièrement intéressante si la famille Π est elle-même paramétrée par un nombre restreint de paramètres comme dans l'exemple 4.3.

La section 4.5 suivante montre une méthode générale de construction de prior conjugués, dans le cadre des *familles exponentielles*, qu'on a introduites au chapitre 2, Section 2.5.

4.5 Familles conjuguées

Dans l'exemple 4.3, nous avons considéré la famille de loi a priori Bêta pour le paramètre θ d'une loi de Bernoulli et montré que pour ce choix la loi a posteriori obtenue est aussi une loi Bêta. On dit que la loi Bêta est une famille de loi *conjuguée* à la loi de Bernoulli. Plus généralement, on parle de loi conjuguée quand pour un modèle Bayésien donné, la loi a posteriori et a priori appartiennent à la même famille de loi. Nous allons voir maintenant un résultat explicitant la sous-famille exponentielle conjuguée à une sous-famille exponentielle donnée.

Supposons que $X = (X_1, \ldots, X_n)$ soit un n-échantillon d'une loi d'une famille exponentielle de dimension d. En notant, comme nous le faisons toujours dans le contexte bayésien, $p(x|\theta)$ pour $p(x;\theta)$, nous avons :

$$p(x|\boldsymbol{\theta}) = \prod_{i=1}^{n} h(x_i) \ \exp\left(\sum_{j=1}^{d} \eta_j(\boldsymbol{\theta}) \sum_{i=1}^{n} T_j(x_i) - nB(\theta)\right) \tag{4.15}$$

où $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$. Posons $\mathbf{t} = (t_1, \ldots, t_{d+1})$ et posons

$$\omega(\mathbf{t}) = \int_{\mathbb{R}^d} \exp\left(\sum_{j=1}^{d} t_j \eta_j(\boldsymbol{\theta}) - t_{d+1} B(\boldsymbol{\theta})\right) d\boldsymbol{\theta}$$
$$\Omega = \left\{\mathbf{t} \in \mathbb{R}^{d+1}, \quad 0 < \omega(\mathbf{t}) < \infty\right\}.$$

Proposition 4.5.1
Supposons que $\Omega \neq \emptyset$. La famille exponentielle $(\pi_{\mathbf{t}}(\boldsymbol{\theta}), \quad \mathbf{t} \in \Omega)$ où,

$$\pi_{\mathbf{t}}(\boldsymbol{\theta}) = \exp\left(\sum_{j=1}^{d} t_j \eta_j(\boldsymbol{\theta}) - t_{d+1} B(\boldsymbol{\theta}) - \log \omega(\mathbf{t})\right)$$

est conjuguée de la famille exponentielle $(p(x|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta)$.

Démonstration. La loi a posteriori est donnée, à une constante de normalisation près par

$$\begin{aligned}\pi(\boldsymbol{\theta}|x) &\propto p(x|\boldsymbol{\theta})\pi_{\mathbf{t}}(\boldsymbol{\theta}) \\ &\propto \exp\left(\sum_{j=1}^{d} \eta_j(\boldsymbol{\theta}) \left(\sum_{i=1}^{n} T_j(x_i) + t_j\right) - (n + t_{d+1}) B(\boldsymbol{\theta})\right) \propto \pi_{\mathbf{s}}(\boldsymbol{\theta}),\end{aligned}$$

où

$$\mathbf{s} = \left(t_1 + \sum_{i=1}^{n} T_1(x_i), \ldots, t_d + \sum_{i=1}^{n} T_d(x_i), n + t_{d+1}\right)^T.$$

■

Exemple 4.4 (Loi conjuguée de la loi gaussienne)**:**
Supposons tout que $(X_1, \ldots, X_n)$ est un n-échantillon d'une v.a. gaussienne $\mathcal{N}(\theta, \sigma_0^2)$, où σ_0 est connu,

$$p(x|\theta) \propto \exp\left(\frac{\theta x}{\sigma_0^2} - \frac{\theta^2}{2\sigma_0^2}\right),$$

qui est une famille exponentielle de dimension 1 avec :

$$T_1(x) = x, \quad \eta_1(\theta) = \frac{\theta}{\sigma_0^2}, \quad B(\theta) = \frac{\theta^2}{2\sigma_0^2}.$$

La famille conjuguée est définie par

$$\pi_{\mathbf{t}}(\theta) = \exp\left(\frac{\theta}{\sigma_0^2}t_1 - \frac{\theta^2}{2\sigma_0^2}t_2 - \log(\omega(t_1, t_2))\right),$$

et donc $\pi_{\mathbf{t}}(\theta) = \mathcal{N}(t_1/t_2, \sigma_0^2/t_2)$, qui est définie pour $(t_1, t_2) \in \Omega = \mathbb{R} \times (\mathbb{R}^+ \setminus \{0\})$. La famille conjuguée est donc $\mathcal{N}(\mu_0, \tau_0)$, où μ_0 peut varier librement et $\tau_0 > 0$. Notons $S = \sum_{i=1}^n X_i$. Pour une telle loi a priori, la moyenne et la variance de la loi a posteriori (qui est gaussienne par construction) sont respectivement données par :

$$\mu(S, n) = \left(\frac{\sigma_0^2}{\tau_0^2} + n\right)^{-1}\left(S + \frac{\eta_0\sigma_0^2}{\tau_0^2}\right), \tag{4.16}$$

et

$$\tau^2(n) = \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}\right)^{-1}. \tag{4.17}$$

Remarquons que, lorsque $n \to \infty$, $\mu(S, n) \approx S/n$ et $\tau^2(n) \approx \sigma_0^2/n$, l'influence de la loi a priori disparaît.

4.6 Risque bayésien, risque intégré

L'approche bayésienne conduit naturellement à un critère global. En effet le paramètre θ est lui-même une réalisation d'un v.a. et P_θ est la distribution conditionnelle de X étant donné $\boldsymbol{\theta} = \theta$. Dans un tel contexte, $R(\theta, \delta) = \mathbb{E}[L(\boldsymbol{\theta}, \delta(X))|\boldsymbol{\theta} = \theta]$, le risque pour la procédure δ lorsque la valeur du paramètre est $\boldsymbol{\theta} = \theta$. Dans le cadre bayésien, il n'y a pas vraiment lieu de s'arrêter à cette étape, car nous pouvons calculer le risque moyen sous la loi a priori du paramètre $\boldsymbol{\theta}$. La quantité intéressante est le *risque bayésien* de δ, que nous notons $r(\delta)$, défini par

$$r(\delta) = \mathbb{E}\left[R(\boldsymbol{\theta}, \delta)\right] = \mathbb{E}\left[L(\boldsymbol{\theta}, \delta(X))\right]. \tag{4.18}$$

Dans le cadre bayésien, une procédure δ est préférable à la procédure δ' si $r(\delta) \leq r(\delta')$. Une procédure δ^* qui minimise le risque bayésien (si une telle procédure existe)

$$r(\delta^*) = \min_\delta r(\delta)$$

est appelée la *procédure de Bayes* ou la *règle de Bayes*. Dans l'exemple précédent, δ_5 est l'unique procédure bayésienne. La méthode consistant à calculer la règle de Bayes en énumérant l'ensemble des règles de décision possibles et en évaluant le risque bayésien pour chaque

i	1	2	3	4	5	6	7	8	9
$r(\delta_i)$	9.6	7.48	8.38	4.92	2.8	3.7	7.02	4.9	5.8
$\max(R(\theta_1,\delta_i), R(\theta_2,\delta_i)$	12	7.6	9.6	5.4	10	6.5	8.4	8.5	6

TABLE 4.1 – Risque bayésien et risque maximal des règles de décision

procédure n'est évidemment pas envisageable dans des situations pratiques. Nous verrons dans un chapitre ultérieur qu'il est possible, pour certaines classes de fonction de perte, de déterminer les estimateurs bayésiens de façon simples, en utilisant les propriétés de l'espérance conditionnelle.

L'approche bayésienne consiste donc à comparer les risques des différentes procédures sur la base de la valeur moyenne sous π de la fonction $\theta \to R(\theta, \delta)$,

$$r(\delta) = \int R(\theta, \delta)\pi(d\theta),$$

où π est la loi a priori du paramètre. Il est possible de considérer ce type de risque moyen même lorsque π n'est pas une probabilité, mais une mesure positive.

Commençons par la description d'un exemple simple.

Exemple 4.5 (Suite de l'exemple 1.11)**:**
Pour illustrer le point de vue bayésien, considérons que dans l'exemple de la prospection pétrolière, un expert pense que la probabilité de trouver du pétrole est de 0.2. Nous pouvons alors traiter le paramètre $\boldsymbol{\theta}$ comme une variable aléatoire, de distribution

$$\pi(\theta_1) = 0.2, \quad \pi(\theta_2) = 0.8.$$

Le risque bayésien de la procédure δ est donc

$$r(\delta) = 0.2R(\theta_1, \delta) + 0.8R(\theta_2, \delta)\ .$$

Dans cet exemple le nombre de décisions non–randomisées possibles est fini (voir la table 1.3). Nous avons maintenant tous les éléments pour décider quelle est la "meilleure" fonction de décision, au sens minimax ou au sens bayésien. D'après la table 4.1, δ_4 est la procédure minimax de risque maximum est 5.4 et δ_5 est la procédure non–randomisée de risque bayésien minimal. Considérons la règle randomisée δ obtenue en choisissant la règle δ_4 ou la règle δ_6 aléatoirement, avec une probabilité $1/2$. Dans ce cas particulier,

$$\frac{1}{2}R(\theta, \delta_4) + \frac{1}{2}R(\theta, \delta_6) = \begin{cases} 4.75 & \text{si} \quad \theta = \theta_1, \\ 4.20 & \text{si} \quad \theta = \theta_2. \end{cases}$$

Le risque maximum est donc 4.75 et est strictement inférieur au risque maximum de la règle de décision δ_4 qui atteint le risque minimax parmi les procédures non-randomisées. D'où l'intérêt des procédures randomisés. Étudions donc plus en détail les procédures randomisées pour cet exemple. Considérons l'ensemble $\mathcal{S}$

$$\mathcal{S} = \{(R(\theta_1, \delta), R(\theta_2, \delta), \delta \in \mathcal{D}^*\}\,,$$

où $\mathcal{D}^*$ est l'ensemble de toutes les procédures de décision (randomisées ou non). Dans le cas présent, l'ensemble $\mathcal{S}$ est l'enveloppe convexe des points $(R(\theta, \delta_1), R(\theta, \delta_2))$, $i = 1, \ldots, 9$,

$$\mathcal{S} = \left\{(r_1, r_2) \ : r_1 = \sum_i \lambda_i R(\theta_1, \delta_i), r_2 = \sum_i \lambda_i R(\theta_2, \delta_i), \lambda_i \geq 0, \sum_i \lambda_i = 1\right\}.$$

Si $\pi(\theta_1) = \gamma = 1 - \pi(\theta_2)$, $0 \leq \gamma \leq 1$, alors l'ensemble des règles de décision ayant un risque bayésien égal à c correspond à l'intersection de $\mathcal{S}$ et des droites d'équation

$$\gamma r_1 + (1-\gamma) r_2 = c. \tag{4.19}$$

En faisant varier le risque c, on obtient ainsi une famille de segments portés par des droites parallèles de pente $-\gamma/(1-\gamma)$. Trouver la règle bayésienne équivaut ici à trouver la plus petite valeur de c pour laquelle l'intersection de la droite (4.19) et de $\mathcal{S}$ est non-vide. Deux cas peuvent se présenter

- l'intersection pour la valeur de c minimale se réduit à un point, l'estimateur bayésien randomisé coïncide avec l'estimateur bayésien non randomisé.
- l'intersection est un segment de droite, auquel cas l'ensemble des points de ce segment sont des estimateurs bayésiens randomisés.

Le changement de loi a priori correspond à changer la pente de la droite $-\gamma/(1-\gamma)$. L'ensemble des règles de décision qui peuvent être des procédures bayésiennes pour certaines lois a priori coïncident avec l'ensemble des segments de pente négatives ou nulles.

Pour trouver les estimateurs minimax randomisés, considérons la famille de carrés,

$$Q(c) = \{(r_1, r_2) : 0 \leq r_1 \leq c, 0 \leq r_2 \leq c\}.$$

Soit c^* la plus petite valeur de c pour laquelle $\mathcal{S} \cap Q(c) \neq \emptyset$. $Q(c^*) \cap \mathcal{S}$ est soit réduit à un point, soit est un segment de droite vertical ou horizontal. Cette intersection coïncide avec l'ensemble des règles de décision minimax randomisées, car s'il existait un estimateur randomisé δ^* tel que $\max(R(\theta_1, \delta^*), R(\theta_2, \delta^*)) = \tilde{c} < c^*$, nous aurions $Q(\tilde{c}) \cap \mathcal{S} \neq \emptyset$, contredisant la définition de c^*. Dans l'exemple considéré, $Q(c^*) \cap \mathcal{S}$ est réduit à un point, appartenant au segment $[\delta_4, \delta_6]$. Une règle de décision δ (randomisée ou non) est dite *inadmissible* s'il existe une règle de décision

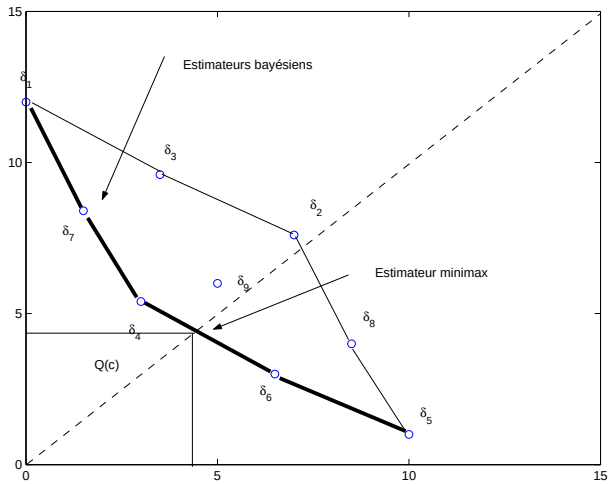


FIGURE 4.2 – Estimateurs bayésiens et minimax

δ' telle que $R(\theta, \delta') \leq R(\theta, \delta)$, pour tout $\theta \in \Theta = \{\theta_1, \theta_2\}$. De façon géométrique, une règle de décision δ de risque $(R(\theta_1, \delta), R(\theta_2, \delta)) = (r_1, r_2)$ est admissible, s'il n'existe pas de point $(x, y) \in \mathcal{S}$ tel que $x \leq r_1$ et $y \leq r_2$, ou de façon équivalente si l'intersection de $\{(x, y) : x \leq r_1, y \leq r_2\}$ et de l'ensemble $\mathcal{S}$ se réduit à (r_1, r_2). La figure montre que les estimateurs

admissibles appartiennent tous à la frontière inférieure de $\mathcal{S}$. (Rappelons que la frontière inférieure d'un ensemble convexe est défini comme l'ensemble des points frontières tels que l'ensemble se situe au-dessus de n'importe quelle tangente à ce point.) Ainsi, l'ensemble des estimateurs admissibles coïncident avec l'ensemble des estimateurs bayésiens.

Chapitre 5

Tests statistiques

Introduction

Un test statistique est un cas particulier de procédure de décision (voir le chapitre 1, section 1.6). Il s'agit de construire une règle de décision (rappelons qu'une telle règle est une fonction des données), permettant de décider si le paramètre θ appartient à telle ou telle région de l'espace Θ. On considère une partition du modèle, $\Theta = \Theta_0 \cup \Theta_1$, et on cherche à déterminer si X est distribuée selon $\theta \in \Theta_0$ ou selon $\theta \in \Theta_1$. Ainsi, un test est une règle de décision à valeurs dans $\{0, 1\}$. Généralement, Θ_0 correspond à une hypothèse « par défaut », que l'on cherche à infirmer ou confirmer au vu des données. Par exemple, on peut se demander si une pièce est biaisée ou non dans un jeu de pile ou face, après avoir observé n lancers. Le modèle est alors $\mathcal{P} = \{\mathcal{B}in(n, \theta) : \theta \in]0, 1[\}$ (le modèle binomial), l'espace des paramètres et $\Theta =]0, 1[$ et l'hypothèse par défaut (absence de biais) correspond au singleton $\Theta_0 = \{0.5\}$, l'hypothèse alternative (présence d'un biais) correspond au reste des possibles, $\Theta_1 =]0, 1[\backslash\{0.5\}$. On voit sur cet exemple que les deux hypothèses ne jouent pas le même rôle : la classe Θ_1 correspondant à l'alternative est beaucoup plus vaste que celle de l'hypothèse par défaut. Ainsi, il sera généralement possible de rejeter l'hypothèse par défaut (aussi appelée « hypothèse nulle ») lorsque par exemple, la moyenne empirique des observations est très éloignée de 0.5. En revanche il sera parfois impossible d'*accepter* l'hypothèse nulle, lorsque l'hypothèse alternative inclut des situations arbitrairement proches de celle-ci : dans l'exemple du pile ou face, l'hypothèse de biais autorise des valeurs de θ arbitrairement proche de 0.5, de sorte qu'on ne peut pas certifier que θ soit exactement égal à 0.5. Les paragraphes suivants formalisent ces idées.

5.1 Tests statistiques et théorie de la décision

5.1.1 Risques et puissance d'un test

Soit $\mathcal{P}$ un modèle statistique, défini sur l'espace des observations $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Soit $\mathcal{P}_0$ et $\mathcal{P}_1$ deux sous-ensembles disjoints tels que $\mathcal{P} = \mathcal{P}_0 \cup \mathcal{P}_1$. Nous disposons d'une observation X et nous nous posons la question de savoir si l'observation X est distribuée sous une loi P_θ où $\theta \in \Theta_0 \subset \Theta$, c'est à dire de *tester* l'hypothèse :

$$H_0 \ : \ \theta \in \Theta_0 \quad \text{contre} \quad H_1 \ : \ \theta \in \Theta_1 \ .$$

L'hypothèse H_0 s'appelle généralement l'*hypothèse de base* ou l'*hypothèse nulle*. L'hypothèse H_1 est appelée *contre-hypothèse* ou *hypothèse alternative*. Une hypothèse est dite *simple* si c'est un singleton, par exemple $\Theta_0 = \{\theta_0\}$. Il est dit *multiple* dans le cas contraire.

Avec le vocabulaire habituel de la théorie de la décision, l'espace des actions est $\mathcal{A} = \{0, 1\}$ (ou $\{\text{acceptation}, \text{rejet}\}$) et une *procédure de test* est une fonction mesurable des observations $\delta : \mathcal{X} \mapsto \mathcal{A} = \{0, 1\}$: si $\delta(x) = 0$, nous acceptons l'hypothèse H_0. Dans le cas contraire, nous rejetons H_0, ou, de façon équivalente, nous acceptons l'hypothèse H_1. Le test statistique définit donc une partition de l'espace des observations $\mathcal{X}$ en deux ensembles mesurables $\mathcal{X}_0 \cup \mathcal{X}_1 = \mathcal{X}$, avec $\mathcal{X}_0 = \{x \in \mathcal{X}, \delta(x) = 0\}$, la région d'*acceptation*. La région $\mathcal{X}_1$ est appelée *région de rejet* ou *région critique*. La fonction de perte utilisée est 0 ou 1 suivant que la décision est correcte ou non. Le risque d'une procédure de test δ est donc donné par

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta\big[\delta(X)\big] = \mathbb{P}_\theta\left[\delta(X) = 1\right] \qquad \forall \theta \in \Theta_0, \qquad (5.1)\\ &= \text{Risque de première espèce} \\ R(\theta, \delta) &= \mathbb{E}_\theta\big[1 - \delta(X)\big] = \mathbb{P}_\theta\left[\delta(X) = 0\right] \qquad \forall \theta \in \Theta_1, \qquad (5.2)\\ &= \text{Risque de deuxième espèce} \end{aligned}$$

Comme on le voit le risque prend deux formes différentes, qu'on appelle respectivement *risque de première espèce* et *risque de deuxième espèce*.

Cette dissymétrie correspond souvent à une réalité pratique : les conséquences de ces deux types d'erreur sont, dans de nombreuses situations, dissymétriques. Ainsi dans l'exemple 1.2, on s'attache en général à contrôler la probabilité que le test réponde « le traitement est efficace » alors qu'il ne l'est pas. Cette probabilité est précisément le risque de première espèce si Θ_0 correspond à l'ensemble des paramètres θ de Θ pour lesquelles $\Delta = 0$. Prenons un autre exemple, si nous testons la présence d'une anomalie sur le système de pilotage d'un avion, décider de façon incorrecte la présence d'une anomalie peut entraîner des coûts financiers ; ne pas la détecter peut avoir des conséquences beaucoup plus dramatiques, sinon encore plus coûteuses. Dans le cas où H_0 est une hypothèse multiple (par exemple, pour le cas du traitement médical, on pourrait envisager l'hypothèse : « le traitement est inefficace ou nocif », soit $\Delta \leq 0$), le risque de première espèce est une fonction de $\theta \in \Theta_0$. Pour s'affranchir de la dépendance en θ du risque de première espèce, on définit le *niveau* comme le risque dans le pire des cas :

Définition 5.1.1. *Le* niveau d'un test δ *est défini comme le pire risque de première espèce*

$$\alpha = \sup_{\theta \in \Theta_0} R(\theta, \delta).$$

C'est-à-dire,

$$\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\delta(X) = 1) = \sup_{\theta \in \Theta_0} \mathbb{E}_\theta(\delta(X)).$$

Le risque de deuxième espèce est la probabilité d'accepter l'hypothèse H_0 alors que cette hypothèse n'est pas vérifiée (lorsque $\theta \in \Theta_1$). Il est courant d'employer la terminologie de « puissance » d'un test plutôt que de son risque de deuxième espèce . Par définition, la puissance est la fonction définie pour $\theta \in \Theta_1$ et δ une procédure de test, par

$$\beta(\theta, \delta) \stackrel{\text{def}}{=} 1 - R(\theta, \delta) = \mathbb{P}_\theta[X \in \mathcal{X}_1] = \mathbb{E}_\theta[\delta(X)]$$

(probabilité d'accepter l'alternative quand celle-ci est vérifiée).

Idéalement, il est souhaitable de disposer d'une procédure de test qui soit telle que les deux risques, première espèce et deuxième espèce, soient simultanément faibles, ou comme il est plutôt d'usage de le présenter, telle que le risque de première espèce est faible et la puissance est forte.

Exemple 5.1 (Pièce biaisée ou non : test de niveau $\alpha \leq 5\%$)**:**
On reprend l'exemple donné en introduction de ce chapitre. Rappelons nous que le modèle est $\Theta = \{\mathcal{B}in(n,p) : p \in]0,1[\}$. L'espace des observations est $X = \{0,1,\ldots,n\}$ et $\Theta =]0,1[$. L'hypothèse nulle est $H_0 : \theta \in \Theta_0$, avec $\Theta_0 = \{0.5\}$ et l'hypothèse alternative est $H_1 : \theta \in \Theta_1$, avec $\Theta_1 =]0,1[\backslash\{0.5\}$.

Puisque l'hypothèse nulle est simple, construire un test de niveau $\alpha \leq 5\%$ (c'est-à-dire, de risque de première espèce égal à $\alpha \leq 0.05$) signifie déterminer une région de rejet $\mathcal{X}_1$ telle que, sous l'hypothèse nulle, la probabilité que X appartienne à $\mathcal{X}_1$ soit inférieure ou égale à $5/100$. Notons $\theta_0 = 0.5$ le paramètre correspondant à l'hypothèse nulle.

Voici un exemple de construction : soient $N_1, N_2 \in \{0,\ldots,n\}$ respectivement les plus grands et plus petits entiers tels que

$$\mathrm{P}_{\theta_0}(\{0,\ldots,N_1\}) \leq 2.5/100 \quad \text{et} \quad \mathrm{P}_{\theta_0}(\{N_2,\ldots,n\} \leq 2.5/100.$$

Par exemple, pour $n = 100$, on obtient $N_1 = 39$ et $N_2 = 61$; avec $\mathrm{P}_{\theta_0}(\{0,\ldots,N_1\}) = \mathrm{P}_{\theta_0}(\{N_1,\ldots,n\} \simeq 1.76\%$. Si l'on définit la région d'acceptation comme étant

$$\mathcal{X}_0 = \{N_1+1,\ldots,N_2-1\},$$

alors la région de rejet est automatiquement $\mathcal{X}_1 = \mathcal{X} \setminus \mathcal{X}_0 = \{0,\ldots,N_1\} \cup \{N_2,\ldots,n\}$. La procédure de test est alors la fonction de décision

$$\delta(X) = \mathbb{1}_{\mathcal{X}_1}(X),$$

et par définition des seuils N_1 et N_2, on a $\mathbb{P}_{\theta_0}(X \in \mathcal{X}_1) \leq 5/100$. Ainsi, le risque de première espèce de la procédure est

$$\alpha = R(\theta_0,\delta) = \mathbb{P}_{\theta_0}(\delta(X)=1) = \mathrm{P}_{\theta_0}(\mathcal{X}_1) \leq 5/100.$$

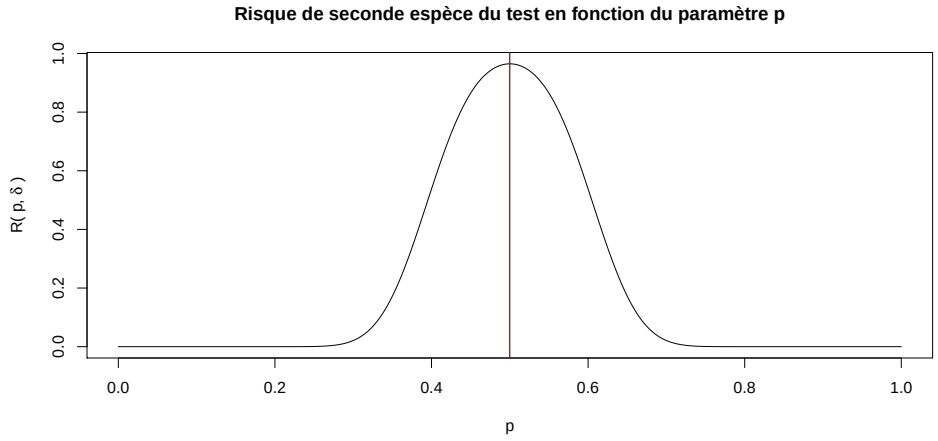
En particulier, pour $n = 100$ et N_1, N_2 comme ci-dessus, on a $\alpha \simeq 2 * 0.0176 \simeq 3.5\%$.

On a ainsi défini une procédure de test telle que, sous l'hypothèse nulle, la probabilité de se tromper (en rejetant l'hypothèse) est inférieure à 5%.

Examinons la *puissance* de notre test. Par définition, c'est la quantité $\beta(\theta) = 1 - R(\theta,\delta)$, pour $\theta \in \Theta_1$. Dans notre exemple,

$$\forall \theta \neq 0.5, \quad \beta(\theta) = 1 - \mathbb{P}_\theta(X \in \mathcal{X}_0) = 1 - \mathrm{P}_\theta\{N_1+1,\ldots,N_2-1\}$$

Ici, l'hypothèse alternative est composite (Θ_1 n'est pas un singleton). La puissance n'est pas une quantité fixée, c'est une fonction de $\theta \in]0,1[(\theta \neq 0.5)$ (qui est inconnue). Sur la figure 5.1, on a tracé le risque de seconde espèce en fonction de $\theta \in]0,1[$, La valeur exclue ($\theta = 0.5$) apparaît comme la bande rouge. On constate graphiquement (et on pourrait le montrer facilement dans ce cas particulier) que la borne supérieure est la limite en $\theta = 0.5$. Comme au chapitre 4, on



pourrait s'affranchir de la dépendance en θ en considérant le risque maximum (sous l'hypothèse alternative),

$$\sup_{\theta\in\Theta_1} R(\theta,\delta) = \lim_{\theta\to 0.5} R(\theta,\delta) = \lim_{\theta\to 0.5} \mathrm{P}_\theta\{N_1+1,\dots,N_2-1\} = \mathrm{P}_{0.5}\{N_1+1,\dots,N_2-1\} = 1-\alpha.$$

Ainsi, sur cet exemple on a $\inf_{\theta\neq 0.5}\beta(\theta) = \alpha$. Plus le niveau du test est contraignant (α petit), plus la puissance du pire des cas est faible (risque de seconde espèce important). Bien sûr, l'égalité $\inf_{\theta\in\Theta_1}\beta(\theta) = \alpha$ n'est pas systématiquement vérifiée dans le cadre des tests, mais cette idée de compromis entre puissance et niveau est à garder en tête pour la suite.

Remarquons pour conclure que l'on a choisi *une* procédure de test arbitraire, parmi toutes celles dont le niveau est inférieur à α. Autrement dit, on aurait pu définir la région d'acceptation de multiples autres manières, par exemple de type $\mathcal{X}_0 = \{0,1,\dots,N_3-1\}$ où $N_3 = \min\{k \leq n : \mathbb{P}_{0.5}(X \leq k) \leq 5/100\}$. Les parties suivantes de ce chapitre développent cette question du choix d'une procédure de test optimale, dans un sens « uniforme » dans certains cas particuliers où la structure du modèle le permet (hypothèses simples ou « monotonie » de la vraisemblance, voir la partie 5.5), ou « en moyenne », dans un cadre bayésien (voir la partie 5.6).

5.1.2 Tests randomisés*

La notion de règle de décision randomisée a été introduite dans un cadre général au chapitre 1, au paragraphe 1.7. Dans le cadre particulier des tests, un test randomisé est caractérisé par une fonction $\phi : \mathcal{X} \to [0,1]$ de la façon suivante :

- Ayant observé X, on simule une variable aléatoire R de loi Bernoulli de paramètre $p = \phi(X)$: ainsi, $R \in \{0,1\}$ avec $\mathbb{P}(R=1\,|\,X) = \mathbb{E}[R\,|\,X] = \phi(X)$.
- Si $R = 1$, nous rejetons l'hypothèse ; autrement, nous acceptons l'hypothèse.

La randomisation de la procédure de test consiste donc à "rajouter" une procédure aléatoire pour "choisir" une hypothèse. En pratique, on n'utilise que très peu (voire pas) les tests randomisés. La notion de randomisation est surtout un artifice mathématique destiné à montrer l'existence de certaines procédures de test optimales, pour tout niveau de risque de première espèce imposé α, en particulier dans le cadre de Neyman-Pearson (que l'on développera dans la partie suivante).

La fonction ϕ est appelée la *fonction critique* du test. Dans le cas d'un test randomisé, l'ensemble $\{x \in \mathcal{X} : \phi(x) = 1\}$ est la *région de rejet* et l'ensemble $\{x \in \mathcal{X} : \phi(x) < 1\}$ est appelée la *région d'acceptation*.

Le risque d'une procédure randomisée δ de fonction critique ϕ s'écrit alors :

$$R(\theta,\delta) = \mathbb{1}_{\Theta_0}(\theta)\mathbb{E}_\theta\big[\phi(X)\big] + \mathbb{1}_{\Theta_1}(\theta)\Big(1 - \mathbb{E}_\theta\big[\phi(X)\big]\Big)\ .$$

Dans la suite, nous utiliserons les expressions suivantes pour les risques de première et deuxième espèce ainsi que pour la puissance d'un test randomisé.

$$\begin{cases} \text{Risque de première espèce :} & \mathbb{E}_\theta\big[\phi(X)\big]\ , & \theta \in \Theta_0 \\ \text{Risque de deuxième espèce :} & 1 - \mathbb{E}_\theta\big[\phi(X)\big]\ , & \theta \in \Theta_1 \\ \text{Puissance :} & \mathbb{E}_\theta\big[\phi(X)\big]\ , & \theta \in \Theta_1\ . \end{cases} \tag{5.3}$$

Si la statistique ϕ prend seulement les deux valeurs 0 et 1, la procédure revient à un test non–randomisé.

5.1.3 Approche de Neyman–Pearson

L'approche proposée par Neyman et Pearson consiste à optimiser la puissance dans une classe donnée de procédures de test. L'idée de cette approche est d'optimiser le risque de deuxième espèce uniformément sous une contrainte de majoration du risque de première espèce. On est dans le cadre de la recherche d'une décision uniformément optimale, sous une contrainte portant sur la famille des décisions considérées (voir le chapitre 4, parties 4.1 et 4.2). Plus précisément on considère une contrainte sur le niveau α du test.

Remarque 5.1.2 (Niveau d'un test randomisé)**.** *Pour un test randomisé de fonction critique ϕ, le niveau s'écrit*

$$\sup_{\theta\in\Theta_0} \mathbb{E}_\theta(\phi(X)).$$

Les niveaux d'un test usuellement utilisés sont $\alpha = 0.1, 0.05, 0.01$, parfois exprimés en pourcentage, $10\%, 5\%, 1\%$. Pour $\alpha \in [0,1]$, on note l'ensemble des tests randomisés de niveau au plus α par $\mathcal{K}_\alpha$,

$$\mathcal{K}_\alpha \overset{\text{def}}{=} \{\delta \ :\ R(\theta,\delta) \le \alpha, \forall\theta \in \Theta_0\} \tag{5.4}$$

Définition 5.1.3. *On dit alors qu'un test $\delta^\star \in \mathcal{K}_\alpha$ est* uniformément plus puissant (U.P.P.) dans la classe $\mathcal{K}_\alpha$*, ou encore* U.P.P. de niveau α *si, pour tout test $\delta \in \mathcal{K}_\alpha$ et pour tout $\theta \in \Theta_1$, le risque de deuxième espèce de $\delta^\star$, $R(\theta,\delta^\star)$ est inférieur au risque de deuxième espèce de δ,*

$$R(\theta,\delta^\star) \le R(\theta,\delta), \qquad \forall\theta \in \Theta_1, \forall\delta \in \mathcal{K}_\alpha.$$

ou, de façon équivalente, si, pour tout $\theta \in \Theta_1$, la puissance du test $\delta^\star$, $\beta(\theta,\delta^\star) = 1 - R(\theta,\delta^\star)$, est supérieure à la puissance du test δ, $\beta(\theta,\delta^\star) \ge \beta(\theta,\delta) = 1 - R(\theta,\delta)$.

La recherche des tests U.P.P. consiste donc à *minimiser uniformément le risque de deuxième espèce* (ou à *maximiser uniformément la puissance*) sous la contrainte que le risque de première espèce est inférieur à un seuil α.

S'il n'existe pas, en général, de test uniformément plus puissant (U.P.P), nous verrons dans la suite de ce chapitre (parties 5.2 et 5.5) qu'il est possible de construire de tels tests dans des cas particuliers importants en pratique.

5.2 Test de Neyman-Pearson (Rapport de vraisemblance) : cas d'hypothèses simples

Un cas où l'on sait construire un test U.P.P. au sens de la définition 5.1.3 est celui des hypothèses simples, où $\Theta_0 = \{\theta_0\}$ et $\Theta_1 = \{\theta_1\}$.

Notons que, dans ce cas, le modèle statistique indexé par $\Theta = \{\theta_0, \theta_1\}$ est toujours dominé, par exemple par la mesure $\nu = \mathrm{P}_{\theta_0} + \mathrm{P}_{\theta_1}$. On notera dans la suite p_0 et p_1 les densités des lois P_{θ_0} et P_{θ_1} par rapport à une mesure de domination ν. La quantité centrale pour construire un test optimal sera la fonction appelée *rapport de vraisemblance*,

$$Z(x) = \frac{p_1(x)}{p_0(x)},\ x \in \mathcal{X}\ . \tag{5.5}$$

Théorème 5.2.1 (Neyman-Pearson I :
caractère U.P.P. du rapport de vraisemblance)
Soient $c > 0$ et $\alpha > 0$ tels que la statistique Z du rapport de vraisemblance vérifie

$$\mathbb{P}_{\theta_0}\left[Z(X) > c\right] = \alpha.$$

Alors,

(1) *Le test*

$$\delta^\star : x \mapsto \mathbb{1}_{Z(x)>c} = \begin{cases} 1 & si\ Z(x) > c \\ 0 & si\ Z(x) \leq c \end{cases} \tag{5.6}$$

est un test uniformément plus puissant de niveau α pour le test de H_0 contre H_1, et sa puissance est supérieure ou égale à α.

(2) *Si $\delta^{\star\star}$ est un autre test U.P.P. de niveau α, alors pour ν-presque tout x,*

$$\delta^{\star\star}(x) = \begin{cases} 1 & si\ Z(x) > c\ , \\ 0 & si\ Z(x) < c\ . \end{cases} \tag{5.7}$$

Remarque 5.2.2. *Le deuxième point nous dit que tout test U.P.P. de niveau α coïncide avec $\delta^\star$, ν-presque partout sur l'ensemble $\{x \in \mathcal{X} : Z(x) \neq c\}$.*

Remarque 5.2.3. *Dans le cas (fréquent) où la fonction de répartition de $Z(X)$ est continue, on obtient c comme un quantile de $Z(X)$, soit $\mathbb{P}_{\theta_0}[Z(X) > c] = \alpha$.*

Démonstration. (Théorème 5.2.1)

Montrons pour commencer que $\delta^\star$ est U.P.P. dans la classe $\mathcal{K}_\alpha$. Il suffit de montrer que pour toute fonction $\phi : \mathcal{X} \to [0,1]$ telle que

$$\mathbb{E}_{\theta_0}\phi(X) \leq \alpha, \tag{5.8}$$

on a

$$\mathbb{E}_{\theta_1}\phi(X) \leq \mathbb{E}_{\theta_1}\delta(X). \tag{5.9}$$

En effet, tout test δ (non randomisé) est une fonction de $\mathcal{X}$ dans $\{0,1\}$, donc est un cas particulier de fonction ϕ comme ci-dessus. De plus, si l'on pose $\phi = \delta$, alors $\delta \in \mathcal{K}(\alpha) \iff$ (5.8) et $\beta(\theta_1, \delta) \leq \beta(\theta_1, \delta^\star) \iff$ (5.9). Dans le cas randomisé (hors-programme, cf. paragraphe 5.1.2), il suffit encore de montrer que (5.8) implique (5.9), car tout test δ est caractérisé

par une *fonction critique* ϕ comme ci-dessus (à valeurs dans $[0,1]$) et l'on a vu que les identités $\mathbb{E}_{\theta_0}\phi(X) = R(\theta_0,\delta)$ (risque de première espèce) et $1-\mathbb{E}_{\theta_1}\phi(X) = R(\theta_1,\delta)$ (risque de seconde espèce) sont vraies dans le cas randomisé comme dans le cas non-randomisé.

Soit donc ϕ une fonction vérifiant (5.8). Si $\delta^\star(x) - \phi(x) > 0$, alors $\delta^\star(x) > 0$ et donc $p_1(x) \geq cp_0(x)$. Si $\delta^\star(x) - \phi(x) < 0$, alors $\delta^\star(x) < 1$ et donc $p_1(x) \leq cp_0(x)$. Dans tous les cas, pour tout $x \in \mathcal{X}$,

$$[\delta^\star(x) - \phi(x)][p_1(x) - cp_0(x)] \geq 0\ . \tag{5.10}$$

Par conséquent, on a

$$\int [\delta^\star(x) - \phi(x)][p_1(x) - cp_0(x)]\nu(dx) \geq 0\ ,$$

ce qui peut se réécrire

$$\int [\delta^\star(x) - \phi(x)]p_1(x)\nu(dx) \geq c\int [\delta^\star(x) - \phi(x)]p_0(x)\nu(dx)\ .$$

Le membre de gauche de l'inégalité précédente est égal à $\mathbb{E}_{\theta_1}\delta^\star(X) - \mathbb{E}_{\theta_1}\phi(X)$ et le membre de droite à $c\{\mathbb{E}_{\theta_0}\delta^\star(X) - \mathbb{E}_{\theta_0}\phi(X)\} = c\{\alpha - \mathbb{E}_{\theta_0}\phi(X)\}$. Si la fonction ϕ vérifie (5.8) alors cette dernière quantité est positive, ce qui prouve (5.9). Ainsi, $\delta^\star$ est U.P.P. dans $\mathcal{K}_\alpha$.

Montrons maintenant que la puissance du test $\delta^\star$, $\beta(\theta_1,\delta^\star) = \mathbb{E}_{\theta_1}\delta^*(X)$, est supérieure ou égale à α. Considérons la fonction critique constante $\phi(x) \equiv \alpha$. Alors $\mathbb{E}_{\theta_1}\phi(X) = \alpha$. De plus, ϕ vérifie la contrainte (5.8) relative au risque de première espèce et on a montré précédemment que ceci implique que $\mathbb{E}_{\theta_1}\delta^\star(X) \geq \mathbb{E}_{\theta_1}\phi(X)$. Ceci montre que $\beta(\theta_1,\delta^\star) \geq \alpha$.

Montrons maintenant le point (2). Soit $\phi^{\star\star}$ la fonction critique d'un test $\delta^{\star\star}$ U.P.P. de niveau α (ici encore, dans le cas non-randomisé, $\phi^{\star\star} = \delta^{\star\star}$, et dans le cas randomisé, par définition de la fonction critique, $\phi^{\star\star} = \mathbb{P}(\delta^{\star\star}(X) = 1|X = x)$).

On a $\mathbb{E}_{\theta_1}\delta^\star(X) - \mathbb{E}_{\theta_1}\phi^{\star\star}(X) = 0$ et $c(\mathbb{E}_{\theta_0}\delta^\star(X) - \mathbb{E}_{\theta_0}\phi^{\star\star}(X)) = c(\alpha - \mathbb{E}_{\theta_0}\phi^{\star\star}(X)) \geq 0$. D'où

$$\int [\delta^\star(x) - \phi^{\star\star}(x)]p_1(x)\nu(dx) \leq c\int [\delta^\star(x) - \phi^{\star\star}(x)]p_0(x)\nu(dx)$$

qui implique

$$\int [\phi^\star(x) - \phi^{\star\star}(x)][p_1(x) - cp_0(x)]\nu(dx) \leq 0\ .$$

Comme par ailleurs (5.10) est valide pour $\phi = \phi^{\star\star}$, on obtient que $\{x:\ [\phi^\star(x) - \phi^{\star\star}(x)][p_1(x) - cp_0(x)] \neq 0\}$ est de mesure ν nulle. D'où le résultat.

■

5.3 Existence d'un test U.P.P. avec randomisation*

Si l'on s'autorise à utiliser des tests randomisés (voir le paragraphe 5.1.2), on a un résultat plus fort, qui garantit l'existence d'un test U.P.P. de niveau α, quel que soit $\alpha \in]0,1[$.

Théorème 5.3.1 (Neyman-Pearson II : existence avec randomisation)
Pour tout $\alpha \in (0,1)$, il existe des constantes $c > 0$ et $\gamma \in [0,1]$, telles que la fonction critique :

$$\phi^\star(x) = \begin{cases} 1 & \text{si } Z(x) > c\ , \\ \gamma & \text{si } Z(x) = c\ , \\ 0 & \text{si } Z(x) < c\ , \end{cases} \tag{5.11}$$

vérifie $\mathbb{E}_{\theta_0}\phi^\star = \alpha$. Le test associé à cette fonction critique est U.P.P. de niveau α. Sa puissance est supérieure ou égale à α. De plus, si $\phi^{\star\star}$ est la fonction critique d'un autre test U.P.P. de niveau α, alors $\phi^{\star\star}$ coïncide avec $\phi^\star$ sur l'ensemble $\{x \in \mathcal{X} : Z(x) \neq c\}$, ν- presque partout.

DÉMONSTRATION. Montrons tout d'abord que l'équation en (c, γ), $0 \leq \gamma \leq 1$:

$$\mathbb{E}_{\theta_0}\phi^\star = \mathbb{P}_{\theta_0}(Z > c) + \gamma \mathbb{P}_{\theta_0}(Z = c) = \alpha \,, \tag{5.12}$$

admet toujours une solution. Remarquons que sous P_{θ_0}, p_0 s'annule avec probabilité nulle et donc Z est une v.a. à valeurs dans $[0, \infty)$. La fonction $c \to \mathbb{P}_{\theta_0}(Z > c)$ est décroissante sur $[0, \infty[$. En tout point $c_0 \in [0, \infty[$, cette fonction est continue à droite et admet des limites à gauche :

$$\lim_{c \downarrow c_0} \mathbb{P}_{\theta_0}(Z > c) = \mathbb{P}_0(Z > c_0) \quad \text{et} \quad \lim_{c \uparrow c_0} \mathbb{P}_{\theta_0}(Z > c) = \mathbb{P}_{\theta_0}(Z \geq c_0).$$

Il existe donc $c_\alpha > 0$ tel que

$$\mathbb{P}_{\theta_0}(Z > c_\alpha) \leq \alpha \leq \mathbb{P}_{\theta_0}(Z \geq c_\alpha).$$

Pour obtenir l'équation (5.12), nous posons $c = c_\alpha$ et

$$\gamma = \begin{cases} 0 & \text{si } \mathbb{P}_{\theta_0}(Z > c_\alpha) = \alpha \\ \frac{\alpha - \mathbb{P}_{\theta_0}(Z > c_\alpha)}{\mathbb{P}_{\theta_0}(Z = c_\alpha)} & \text{si } \mathbb{P}_{\theta_0}(Z > c_\alpha) < \alpha \,. \end{cases}$$

Dans le second cas, on a bien $\gamma \in [0, 1]$ car $\mathbb{P}_{\theta_0}(Z = c_\alpha) = \mathbb{P}_{\theta_0}(Z \geq c_\alpha) - \mathbb{P}_{\theta_0}(Z > c_\alpha) \geq \alpha - \mathbb{P}_{\theta_0}(Z > c_\alpha) > 0$. Le test $\delta^\star$ de fonction critique $\phi^\star$ définie par (5.11) avec $c = c_\alpha$ est donc un test de niveau α.

Le reste de l'énoncé du théorème (affirmant que $\delta^\star$ est U.P.P. de niveau α, que sa puissance est $\geq \alpha$, et que tout autre test U.P.P. de niveau α coïncide presque partout avec ϕ^* en dehors de l'ensemble $\{Z(x) = c\}$ se montre exactement comme dans la preuve du théorème 5.2.1.

■

Remarquons que si la loi de $Z(X)$ n'a pas d'atomes sous $\mathbb{P}_{\theta_0}$, c'est-à-dire si $\mathbb{P}_{\theta_0}(Z(X) = c) = 0$ pour tout $c \geq 0$, on peut choisir $\gamma = 0$ dans (5.11) et donc obtenir un test U.P.P. non–randomisé.

5.4 Exemples

Exemple 5.2 (Deux variables gaussiennes scalaires)**:**
Supposons que $p_i(x) = 1/\sqrt{2\pi\sigma_i^2}\exp(-(x-\mu_i)^2/2\sigma_i^2)$, $i = 0, 1$ sont les densités de probabilité de deux variables gaussiennes scalaires de moyenne et de variance (μ_0, σ_0^2) et (μ_1, σ_1^2), respectivement, avec $\mu_0 < \mu_1$. Le rapport de vraisemblance est alors donné par :

$$Z(x) = \frac{\sigma_0}{\sigma_1} \exp\left(-\frac{1}{2\sigma_1^2}(x - \mu_1)^2 + \frac{1}{2\sigma_0^2}(x - \mu_0)^2 \right). \tag{5.13}$$

Considérons d'abord le cas où $\sigma_0 = \sigma_1$. Alors, les termes d'ordre 2 en x se compensent dans l'expression ci-dessus, et le rapport de vraisemblance s'écrit $Z(x) = C\exp(\frac{x(\mu_1 - \mu_0)}{2\sigma_0^2})$ (où C est une constante), qui est une fonction croissante de x (dans le cas où $\mu_1 > \mu_0$). Un test de rapport de vraisemblance de type (5.6) aura donc une région critique de la forme

$$\mathcal{X}_1 = \{x \ : \ x > K\}\,. \tag{5.14}$$

Il reste à déterminer K, étant donné un niveau de test α souhaité. Pour cela, il suffit de choisir K tel que

$$\mathcal{N}_{(\mu_0, \sigma_0^2)}[K, +\infty) = \alpha.$$

Cette équation en K admet une unique solution pour tout $\alpha \in]0,1[$ car la fonction de répartition de la loi Gaussienne (continue et strictement croissante sur $\mathbb{R}$) est une bijection de $]0,1[$ dans $\mathbb{R}$. La solution K est le quantile q_{α,μ,σ^2} de la loi Gaussienne (μ, σ^2). Ces quantiles sont tabulés dans n'importe quel logiciel de calcul numérique (`R`, `python`, `Matlab`, ...). Dans la figure 5.1, nous avons représenté les régions critiques du test lorsque $(\mu_0, \sigma_0^2) = (-1, 1)$ et $(\mu_1, \sigma_1^2) = (1, 1)$, la variance est identique sous les deux alternatives. Insistons sur le fait que dans ce cas particulier, le rapport de vraisemblance est une fonction monotone croissante de x, ce qui simplifie le calcul de la région critique. En particulier, on voit dans cet exemple que la région critique ne dépend pas de l'alternative (le paramètre μ_1).

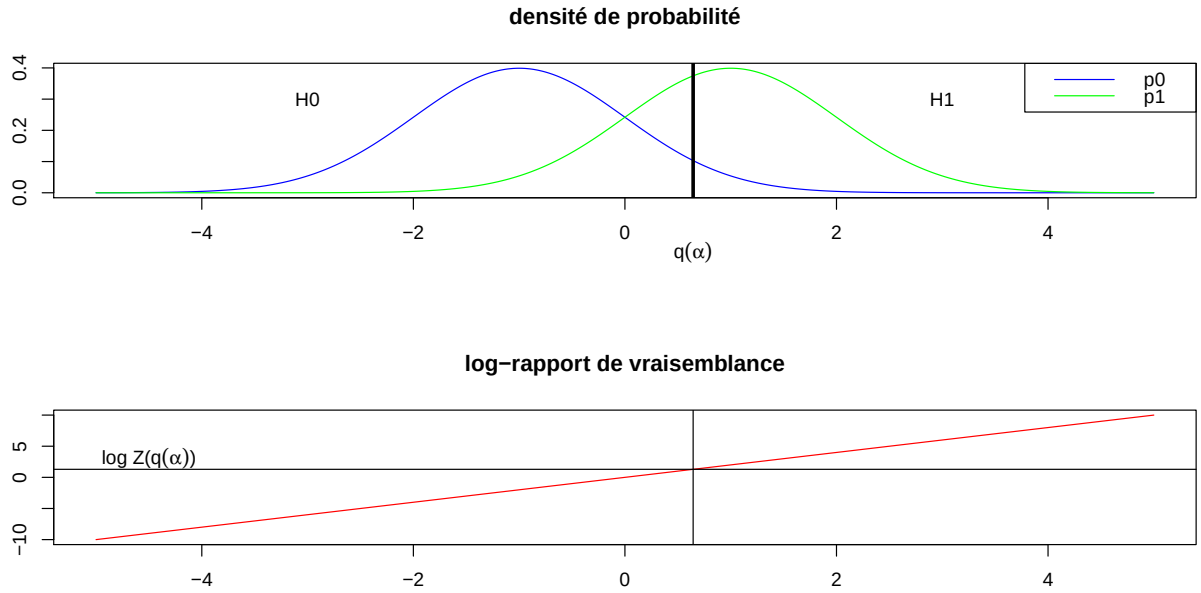


FIGURE 5.1 – Panneau du haut : densité de probabilité de deux v.a. gaussiennes de moyenne et de variance $(\mu_0, \sigma_0^2) = (-1, 1)$ et $(\mu_1, \sigma_1^2) = (1, 1)$. Panneau du bas : rapport de vraisemblance $Z(x) = p_1(x)/p_0(x)$.

Passons au cas général, c'est-à-dire ne supposons plus que $\sigma_0 = \sigma_1$. Au vu de l'expression (5.13) du rapport de vraisemblance, la région critique d'un test de Neyman-Pearson de type (5.6) sera de la forme

$$\mathcal{X}_1 = \left\{ x \ : \ -\frac{1}{2\sigma_1^2}(x-\mu_1)^2 + \frac{1}{2\sigma_0^2}(x-\mu_0)^2 > C \right\}, \tag{5.15}$$

avec $C = \log(c) + (1/2)\log(\sigma_1^2/\sigma_0^2)$. La région critique est donc délimitée par les racines d'une équation du second degré en x. Remarquons qu'il est toujours possible de fixer C tel que le polynôme ait deux racines distinctes.$(x_1(C), x_2(C))$. Lorsque $\sigma_0 < \sigma_1$, le terme dominant du polynôme est positif et la région critique se trouve à l'extérieur des racines x_1, x_2. Si au contraire, $\sigma_0 > \sigma_1$, s'est l'intervalle entre les deux racines. Cependant, la détermination explicite de C pour un niveau de test α souhaité est plus délicat, car le risque de première espèce pour un test de ce type est donné par $\mathcal{N}_{(\mu_0,\sigma_0)}[x_1(C), x_2(C)]$ (lorsque $\sigma_0 > \sigma_1$) ou $1 - \mathcal{N}_{(\mu_0,\sigma_0)}[x_1, x_2]$ (lorsque $\sigma_0 < \sigma_1$). Inverser cette relation (pour obtenir C en fonction du risque souhaité) nécessite un

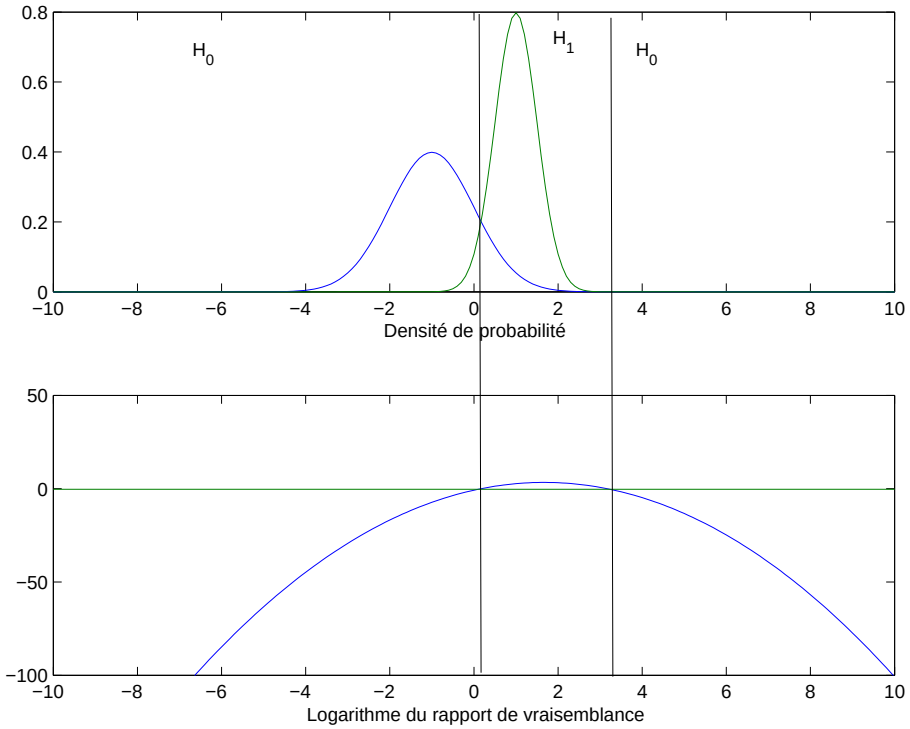


FIGURE 5.2 – Panneau du haut : densité de probabilité de deux v.a. gaussiennes de moyenne et de variance $(\mu_0, \sigma_0^2) = (-1, 1)$ et $(\mu_1, \sigma_1^2) = (1, 0.5)$. Panneau du bas : rapport de vraisemblance $Z(x) = p_1(x)/p_0(x)$.

recours à des méthodes numériques.

Nous avons représenté dans la figure 5.2 des régions critiques de ce test lorsque $(\mu_0, \sigma_0^2) = (-1, 1)$ et $(\mu_1, \sigma_1^2) = (1, 0.5)$, en fixant arbitrairement le seuil c dans (5.6) à 1.

Exemple 5.3 (Test de la moyenne de v.a. gaussiennes : variance connue)**:**
Soit $(X_1, \ldots, X_n)$ des v.a. gaussiennes indépendantes $\mathcal{N}(\mu_i, \sigma^2)$ où σ^2 est supposé connu. Considérons l'hypothèse de base $H_0 = \{\mu_i = 0, i = 1, \ldots, n\}$ et l'hypothèse alternative $H_1 = \{\mu_i = \mu, i = 1, \ldots, n\}$ où μ est une constante connue. Nous cherchons à déterminer le test Neyman-Pearson de niveau α. Il s'agit ici d'un test d'hypothèse simple classique. Formons le rapport de vraisemblance,

$$\begin{aligned} Z(\mathbf{x}) = \frac{p_1(x_1, \ldots, x_n)}{p_0(x_1, \ldots, x_n)} &= \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 \right), \\ &= \exp\left(\frac{n\mu}{\sigma^2} \bar{x} - \frac{n\mu^2}{2\sigma^2} \right), \end{aligned}$$

où $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$ et donc $Z(\mathbf{x}) = \tilde{Z}(\bar{x})$ dépend uniquement de la statistique exhaustive $\bar{x}$. On remarque que la fonction $\bar{x} \to \tilde{Z}(\bar{x})$ est une fonction strictement monotone de $\bar{x}$, croissante si $\mu > 0$ et décroissante dans le cas contraire. Si $\mu \geq 0$, la condition $\tilde{Z}(\bar{x}) \geq c$ est équivalente à $\bar{x} \geq d$. Pour déterminer le seuil d, nous devons résoudre l'équation :

$$\mathbb{P}_{\theta_0}(\bar{X} \geq d) = \alpha. \tag{5.16}$$

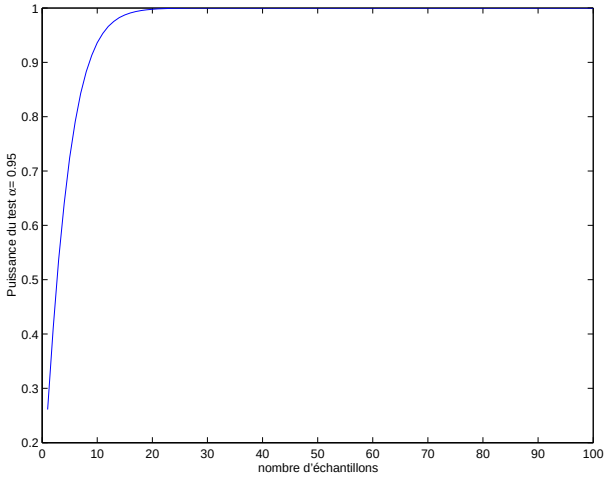


FIGURE 5.3 – Puissance du test U.P.P de $H_0 = \{\mu = 0\}$ contre $H_1 = \{\mu = 1\}$ de niveau $\alpha = 0.95$ en fonction de la taille de l'échantillon.

Comme, sous P_{θ_0}, la variable aléatoire $\sqrt{n}\bar{X}/\sigma$ est distribuée suivant une loi $\mathcal{N}(0,1)$. Notons $z(\alpha)$ le quantile d'ordre α de la loi gaussienne standard,

$$\Phi(z(\alpha)) = \int_{-\infty}^{z(\alpha)} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)dx = \alpha.$$

L'équation (5.16) admet comme seule solution $d = z(1-\alpha)\sigma/\sqrt{n}$. Il est intéressant de remarquer que le test ne dépend pas de μ, la valeur de la moyenne sous l'alternative. La puissance du test est alors donnée par :

$$\mathbb{P}_\mu(\bar{X} \geq z(1-\alpha)\sigma/\sqrt{n}) = 1 - \Phi(z(1-\alpha) - \sqrt{n}\mu/\sigma).$$

Nous avons représenté dans la figure 5.3 la fonction puissance dans le cas particulier où $\mu = 1$, $\sigma = 1$ et $\alpha = 0.05$ $(z(1-\alpha) = 1.6449)$, pour des tailles d'échantillon variant de 10 à 1000. Ce test se généralise aisément au cas où la moyenne sous la contre-alternative n'est pas constante $H_1 = \{\mu_i = \nu_i, i = 1, \ldots, n\}$. Dans ce cas particulier, le rapport de vraisemblance est de la forme :

$$Z(\mathbf{x}) = \frac{p_1(x_1, \ldots, x_n)}{p_0(x_1, \ldots, x_n)} = \exp\left(\frac{1}{\sigma^2}\sum_{i=1}^n \nu_i X_i - \frac{1}{2\sigma^2}\sum_{i=1}^n \nu_i^2\right).$$

Le rapport de vraisemblance est cette fois fonction de la statistique $\sum_{i=1}^n \nu_i X_i$, et le test de rapport de vraisemblance est alors de la forme :

$$\sum_{i=1}^n \nu_i X_i \geq d.$$

En remarquant que sous l'hypothèse de base, $\sum_{i=1}^n \nu_i X_i/\sigma\sqrt{\sum_{i=1}^n \nu_i^2}$ est une loi gaussienne standard, on obtient un test de niveau α en rejetant l'hypothèse de base si :

$$\sum_{i=1}^n \nu_i X_i \geq z(1-\alpha)\sigma\sqrt{\sum_{i=1}^n \nu_i^2}.$$

Ce test est à la base de nombreuses applications en communication numérique et en théorie du signal radar.

Exemple 5.4 (Variance d'une gaussienne : moyenne connue)**:**
Soit $(X_1, \ldots, X_n)$ un n-échantillon d'une v.a. gaussienne $\mathcal{N}(0, \theta)$. Nous souhaitons tester l'hypothèse $\theta = \theta_0$ contre $\theta = \theta_1$, où $0 < \theta_0 < \theta_1$. Le rapport de vraisemblance est de la forme :

$$Z(x_1, \ldots, x_n) = \left(\frac{\theta_0}{\theta_1}\right)^{n/2} \exp\left(-\left(\frac{1}{2\theta_1} - \frac{1}{2\theta_0}\right)\sum_{i=1}^{n} x_i^2\right).$$

L'événement $Z(x_1, \ldots, x_n) > c$ est équivalent à $\sum_{i=1}^n x_i^2 > d$ pour un d convenablement choisi (c'est un cas particulier de rapport de vraisemblance monotone, que nous étudierons plus en détail dans la suite). Pour déterminer le seuil d, nous devons donc résoudre l'équation :

$$\mathrm{P}_{\theta_0}\left(\sum_{i=1}^{n} X_i^2 \geq d\right) = \alpha,$$

Comme $\sum_{i=1}^n X_i^2/\theta_0$ est distribué suivant une loi du χ^2 centre à n degré de liberté, on peut déterminer d à partir des quantiles de cette loi.

Exemple 5.5 (Un cas de loi discrète)**:**
Soient $(X_1, \ldots, X_n)$ n variables i.i.d. de loi de Bernoulli de paramètre θ. On suppose que $H_0 = \{\theta = \theta_0\}$ et $H_1 = \{\theta = \theta_1\}$, où $0 < \theta_0 < \theta_1$. En posant $S(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$ et $Z(s; \theta_0, \theta_1) = (\theta_1/\theta_0)^s((1-\theta_1)/(1-\theta_0))^{n-s}$, le Théorème de Neyman–Pearson implique que le test de fonction critique

$$\phi^\star(X_1, \ldots, X_n) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{si } Z(S(X_1, \ldots, X_n); \theta_0, \theta_1) > c \\ \gamma & \text{si } Z(S(X_1, \ldots, X_n); \theta_0, \theta_1) = c \\ 0 & \text{si } Z(S(X_1, \ldots, X_n); \theta_0, \theta_1) < c \end{cases}$$

est U.P.P. dans la classe des tests de niveau α. La fonction $s \mapsto Z(s; \theta_0, \theta_1)$ est monotone en s, ce qui implique que le test précédent peut s'écrire

$$\phi^\star(X_1, \ldots, X_n) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{si } S(X_1, \ldots, X_n) > m \\ \gamma & \text{si } S(X_1, \ldots, X_n) = m \\ 0 & \text{si } S(X_1, \ldots, X_n) < c \end{cases}$$

où $m \in \mathbb{N}$ et γ sont des constantes telles que

$$\alpha = \mathbb{E}_{\theta_0}\phi^\star = \mathbb{P}_{\theta_0}(S > m) + \gamma \mathbb{P}_{\theta_0}(S = m)\,.$$

Comme $S(X_1, \ldots, X_n)$ est distribué suivant une loi binomial de paramètre θ sous P_{θ_0}, nous pouvons déterminer m et γ en résolvant

$$\alpha = \sum_{j=m+1}^{n} \binom{n}{j} \theta_0^j (1-\theta_0)^{n-j} + \gamma \binom{n}{m} \theta_0^m (1-\theta_0)^{n-m}\,.$$

Sauf pour les valeurs de α telles que

$$\alpha = \sum_{j=m+1}^{n} \binom{n}{j} \theta_0^j (1-\theta_0)^{n-j}\,,$$

pour un entier m (auquel cas nous pouvons poser $\gamma = 0$), le test U.P.P. est un test randomisé.

5.5 Rapport de vraisemblance monotone

Considérons un modèle statistique $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta \subset \mathbb{R}\}$ paramétrique de paramètre scalaire, $\Theta \subseteq \mathbb{R}$. La situation la plus simple, quand on cherche à généraliser les tests au delà des tests d'hypothèses simples est de supposer que le paramètre inconnu est scalaire et que l'hypothèse de base est *unilatérale* : $H_0 = \{\theta \leq \theta_0\}$, où θ_0 est un paramètre donné. De façon générale, le test le plus puissant de l'hypothèse H_0 contre l'alternative $\{\theta = \theta_1\}$, avec $\theta_1 > \theta_0$ dépend de la valeur de θ_1, et on ne sait pas construire de test uniformément plus puissant de l'hypothèse H_0 contre l'alternative $H_1 = \{\theta > \theta_0\}$. Nous allons voir toutefois qu'il existe des tests U.P.P. lorsque l'on impose une hypothèse supplémentaire sur la structure statistique du modèle. Nous utiliserons l'hypothèse suivante dans ce paragraphe :

(MON) le modèle statistique $\mathcal{P}$ est dominé, $\mathrm{P}_\theta(dx) = p_\theta(x)\mu(dx)$, et il existe une statistique scalaire $T(X)$ telle que pour tout θ et θ' tels que $\theta < \theta'$, le rapport de vraisemblance $Z_{\theta,\theta'}(x) = \frac{p_{\theta'}(x)}{p_\theta(x)}$ est une fonction strictement croissante de $T(x)$ sur son ensemble de définition, c'est-à-dire il existe une fonction $\tilde{Z}_{\theta,\theta'} : \mathbb{R} \to \mathbb{R}$, strictement croissante, telle que

$$\frac{p_{\theta'}(x)}{p_\theta(x)} = \tilde{Z}_{\theta,\theta'}(T(x)),$$

pour tout x tel que $p_\theta(x) > 0$ ou $p_{\theta'}(x) > 0$.

Exemple 5.6 (Loi gaussienne)**:**
les familles gaussiennes $\mathcal{N}(\theta, 1)$ $(\Theta = \mathbb{R})$ et $\mathcal{N}(0, \theta^2)$ $(\Theta = \mathbb{R}^+)$ sont des exemples pour lesquels les rapports de vraisemblance sont monotones, puisque l'on a, dans ces cas respectifs :

$$\frac{p_{\theta'}(x)}{p_\theta(x)} = \exp\left\{(\theta' - \theta)n\bar{x} - (n/2)(\theta'^2 - \theta^2)\right\},$$
$$\frac{p_{\theta'}(x)}{p_\theta(x)} = \sqrt{\frac{\theta^2}{\theta'^2}} \exp\left\{-\frac{1}{2}(\theta'^{-2} - \theta^{-2}) \sum_{i=1}^n x_i^2\right\}.$$

où $x = (x_1, \ldots, x_n)$.

Exemple 5.7 (Loi binomiale)**:**
Soit $X_1, \ldots, X_n$ un n-échantillon d'une loi de Bernoulli $\mathcal{B}er(\theta)$. Nous avons pour $\theta, \theta' \in [0, 1]$,

$$\frac{p(x_1, \ldots, x_n; \theta')}{p(x_1, \ldots, x_n; \theta)} = \frac{(1 - \theta')^n}{(1 - \theta)^n} \left(\frac{\theta'(1 - \theta)}{\theta(1 - \theta')}\right)^s,$$

où $s = \sum_{i=1}^n x_i$ et le rapport de vraisemblance est strictement monotone par rapport à s.

De façon générale, si l'observation $(X_1, \ldots, X_n)$ est un n-échantillon i.i.d. d'une famille exponentielle de densité associée à la paire (h, T), où T est une statistique scalaire :

$$p(x; \theta) = h(x) \exp(\phi(\theta)T(x) - \psi(\theta)),$$

le rapport de vraisemblance est monotone si la fonction $\theta \to \phi(\theta)$ est monotone. Si $\theta \to \phi(\theta)$ est croissante alors le rapport de vraisemblance est une fonction croissante de T. Il est décroissant dans le cas contraire.

Remarquons que l'hypothèse *(MON)* implique que pour tout $\theta < \theta'$ et tout d, la condition $p_{\theta'}(x)/p_\theta(x) \geq d$ s'écrit de manière équivalente $T(x) \geq c(\theta, \theta', d)$.

Lemme 5.5.1
Supposons que θ est un paramètre scalaire et (MON) *est vérifiée. Soit $\varphi : \mathbb{R} \to \mathbb{R}$ une fonction monotone croissante (au sens large). Alors $\theta \mapsto g(\theta) = \mathbb{E}_\theta \varphi \circ T(X)$ est une fonction croissante (au sens large).*

Démonstration. Soit $\theta_1 < \theta_2$, $A \stackrel{\text{def}}{=} \{x : p_{\theta_1}(x) > p_{\theta_2}(x)\}$, $a \stackrel{\text{def}}{=} \sup_{x \in A} \varphi \circ T(x)$, $B \stackrel{\text{def}}{=} \{x : p_{\theta_1}(x) \leq p_{\theta_2}(x)\}$, et $b \stackrel{\text{def}}{=} \inf_{x \in B} \varphi \circ T(x)$. Sous *(MON)*, le rapport de vraisemblance $p_{\theta_2}(x)/p_{\theta_1}(x)$ est une fonction monotone croissante de $T(x)$. Par hypothèse, la fonction φ est monotone croissante ; par conséquent, si $x \in A$ et $y \in B$, $\varphi \circ T(y) \geq \varphi \circ T(x)$, d'où $b \geq a$. Par conséquent,

$$\begin{aligned} g(\theta_2) - g(\theta_1) &= \int \varphi \circ T(x) \{p_{\theta_2}(x) - p_{\theta_1}(x)\} \mu(dx) \\ &\geq a \int_A \{p_{\theta_2}(x) - p_{\theta_1}(x)\} \mu(dx) + b \int_B \{p_{\theta_2}(x) - p_{\theta_1}(x)\} \mu(dx) \\ &= (b-a) \int_B \{p_{\theta_2}(x) - p_{\theta_1}(x)\} \mu(dx) \geq 0 \, . \end{aligned}$$

■

Théorème 5.5.2
Supposons que l'hypothèse (MON) *est vérifiée. Soit $\theta_0 \in \Theta$, $\alpha \in (0,1)$ et c tels que*

$$\mathbb{P}_{\theta_0}(T(X) > c) = \alpha. \tag{5.17}$$

Alors :

1. *Le test de l'hypothèse de base $H_0 = \{\theta = \theta_0\}$ contre l'alternative $H_1 = \{\theta > \theta_0\}$ défini par :*
$$\delta(x) = \begin{cases} 1 & \text{si } T(x) > c, \\ 0 & \text{si } T(x) \leq c, \end{cases} \tag{5.18}$$
est uniformément plus puissant de niveau α.
2. *La fonction $\theta \mapsto g(\theta) = \mathbb{E}_\theta \delta(X)$ est croissante sur l'ensemble $\{\theta \in \Theta : \beta(\theta, \delta) < 1\}$.*
3. *Le test δ donné par (5.18) est également U.P.P. pour l'hypothèse de base $H_0 = \{\theta \leq \theta_0\}$ contre l'alternative $H_1 = \{\theta > \theta_0\}$ au niveau α (*i.e.*dans la classe $\mathcal{K}_\alpha$).*

Démonstration. (a) On montre pour commencer que δ est uniformément plus puissant pour le test de $H_0 : \{\theta = \theta = \theta_0\}$ contre $H_1 : \{\theta > \theta_0\}$. Soit $\theta_1 > \theta_0$ et considérons tout d'abord les hypothèses $H_0 = \{\theta = \theta_0\}$ contre $H_1 = \{\theta = \theta_1\}$. D'après l'hypothèse (MON), le rapport de vraisemblance s'écrit $Z_{\theta_0,\theta_1}(x) \stackrel{\text{def}}{=} \frac{p_1(x)}{p_0(x)} = \tilde{Z}_{\theta_0,\theta_1}(T(x))$. Comme la fonction $t \mapsto \tilde{Z}_{\theta_0,\theta_1}(t)$ est strictement croissante, la condition $\{T(x) > c\}$ est équivalente à la condition $\{Z_{\theta_0,\theta_1}(x) > d_{\theta_0,\theta_1}\}$, avec $d_{\theta_0,\theta_1} = \tilde{Z}_{\theta_0,\theta_1}(c)$. Ainsi, la définition de δ dans (5.18) est équivalente à

$$\delta(x) = \begin{cases} 1 & \text{si } Z_{\theta_0,\theta_1}(x) > d_{\theta_0,\theta_1}, \\ 0 & \text{si } Z_{\theta_0,\theta_1}(x) \leq d_{\theta_0,\theta_1} \end{cases}$$

Le test δ est donc un test de rapport de vraisemblance de type (5.6). De plus, l'hypothèse (5.17) portant sur la puissance du test se réécrit sous la forme :

$$\mathbb{P}_{\theta_0}\left(Z_{\theta_0,\theta_1} > d_{\theta_0,\theta_1}\right) = \alpha. \tag{5.19}$$

Les hypothèses du théorème de Neyman-Pearson (5.2.1) sont donc satisfaites, de sorte que le test δ est U.P.P. de niveau α pour le test de H_0 contre H_1. Comme la fonction de test δ a été construite indépendemment de θ_1 (on a seulement supposé (5.17), qui ne fait pas intervenir θ_1), le raisonnement est valide pour tout $\theta_1 > \theta_0$. Ainsi, en considérant maintenant le test hypothèses $H_0 : \{\theta = \theta_0\}$ contre $H_1 : \{\theta > \theta_0\}$, on a

(1) La majoration sur le risque du première espèce : $R(\theta_0, \delta) = \mathbb{E}_{\theta_0}(\delta(X)) = \alpha \leq \alpha$

(2) D'après Neyman-Pearson et le raisonnement ci-dessus, pour tout autre test δ' tel que $R(\theta_0, \delta') \leq \alpha$, pour tout $\theta_1 > \theta$, le fait que $R(\theta_1, \delta') \geq R(\theta_1, \delta)$.

Ces deux conditions montrent que δ est U.P.P de niveau α pour le test de $H_0 : \{\theta = \theta_0\}$ contre $H_1 : \{\theta > \theta_0\}$.

(b) Comme $\delta(x)$ (définie par (5.18)) est une fonction monotone croissante de $T(x)$, $\delta(x) = \mathbb{1}_{[c,\infty[}(T(x))$, le Lemme 5.5.1 montre que la fonction $\theta \mapsto g(\theta) = \mathbb{E}_\theta(\delta(X)) = \mathbb{E}_\theta(\mathbb{1}_{[c,\infty[} \circ T(X))$ est croissante.

(c) D'après le point précédent, pour tout $\theta < \theta_0$, le risque de première espèce de δ pour θ est $R(\theta, \delta) = \mathbb{E}_\theta(\delta(X)) = g(\theta) \leq g(\theta_0) = \alpha$. Ainsi, δ est aussi un test de niveau α pour l'hypothèse $H_0 = \{\theta : \theta \leq \theta_0\}$. Pour montrer qu'il est uniformément plus puissant, considérons un autre test δ' de niveau α pour l'hypothèse $H_0 = \{\theta : \theta \leq \theta_0\}$. Il s'agit de montrer que pour tout $\theta_1 > \theta_0$, $R(\theta_1, \delta') > R(\theta_1, \delta)$. Par hypothèse, δ' satisfait

$$\forall \theta \leq \theta_0, \quad R(\theta, \delta') \leq \alpha.$$

Ceci vaut en particulier pour $\theta = \theta_0$, donc δ' appartient à la classe

$$\mathcal{K}_\alpha = \{\phi : \mathcal{X} \to \{0, 1\} \ : \ R(\theta_0, \phi) \leq \alpha\}.$$

Puisque l'on a montré au (a) que δ est U.P.P. dans la classe $\mathcal{K}_\alpha$ pour le test de $\{\theta = \theta_0\}$ contre l'alternative $H_1 : \{\theta > \theta_0\}$, on a bien, pour tout $\theta_1 > \theta_0$, $R(\theta_1, \delta') > R(\theta_1, \delta)$, ce qu'il fallait démontrer.

■

Remarque 5.5.3 (Existence pour tout α en autorisant la randomisation*)**.** *De même que dans le cas de tests d'hypothèses simples (voir le paragraphe 5.3), si l'on s'autorise à randomiser la procédure de test, on peut montrer l'existence d'un test U.P.P. de niveau α basé sur le rapport de vraisemblance, pour tout $\alpha \in]0, 1[$. La fonction critique du test sera alors de type*

$$\phi(x) = \begin{cases} 1 & si\ T(x) > c, \\ \gamma & si\ T(x) = c, \\ 0 & si\ T(x) < c, \end{cases} \tag{5.20}$$

où les constantes c et γ sont solutions de l'équation :

$$\mathbb{E}_{\theta_0}\phi(X) = \mathbb{P}_{\theta_0}(T(X) > c) + \gamma \mathbb{P}_{\theta_0}(T(X) = c) = \alpha \ .$$

L'argument est le même que dans la preuve du théorème 5.5.2, à ceci près qu'il faut faire appel au théorème d'existence 5.3.1 à la place du théorème 5.2.1.

Remarque 5.5.4 (Sens des inégalités)**.** *Le choix du sens des inégalités dans l'argument de cette partie est arbitraire, les résultats restent valides dans lorsque le rapport de vraisemblance est une fonction décroissante de $T(x)$ et/ou lorsque l'hypothèse nulle est de type $H_0 : \{\theta \leq \theta_0\}$, à condition d'inverser le sens des inégalités dans la définition du test (5.18). En effet, dans le premier cas, le rapport de vraisemblance est alors une fonction croissante de $-T(x)$ et dans le deuxième cas, on peut re-paramétrer le modèle en posant $\tau = -\theta$ et l'hypothèse nulle s'écrit $H_0 : \{\tau \leq \tau_0\}$.*

Exemple 5.8 (Modèle binomial – suite de l'exemple 5.7)**:**
Considérons une observation S d'un modèle binomial $\mathrm{P}_\theta = B(n, \theta)$,

$$\mathrm{P}_\theta(\{s\}) = \binom{n}{s}\theta^s(1-\theta)^{n-s} = \binom{n}{s}\exp(s\log(\theta/1-\theta) + n\log(1-\theta)).$$

Considérons l'hypothèse de base $H_0 = \{\theta \geq \theta_0\}$. Cet type de problème s'introduit naturellement dans le cadre de problème de contrôle de qualité. On inspecte la qualité d'un lot d'objets manufacturés par sondage. On tire ainsi un échantillon de taille n (avec replacement) ; chaque objet a une probabilité θ d'être défectueux. Le rapport de vraisemblance

$$Z_{\theta,\theta'}(s) = \Big(\frac{1-\theta'}{1-\theta}\Big)^n \Big(\frac{\theta'(1-\theta)}{\theta(1-\theta')}\Big)^s$$

est monotone par rapport à s. Si $\theta' < \theta$, il est strictement décroissant. Ainsi, pour tout (α, c) tel que $\mathbb{P}_{\theta_0}(S < c) = \alpha$, le test δ qui consiste à rejeter H_0 si S est inférieure à c est uniformément plus puissant au niveau α.

Une autre façon de procéder est de tirer dans l'échantillon jusqu'à trouver exactement m objets défectueux. Notons $T_0 = 0$ et définissons récursivement les instants $T_i = \inf\{k > T_{i-1}, X_k = 1\}$, c'est-à-dire l'instant où l'on tire le i-ième objet défectueux. On établit aisément que $Y_i = T_i - T_{i-1}$ suit une loi géométrique :

$$\mathbb{P}_\theta[Y_i = y_i] = \theta(1-\theta)^{y_i},$$

et que les variables $Y_1, Y_2, \ldots, Y_m$ sont indépendantes. La loi jointe de ces observations est donc donnée par :

$$\begin{aligned}\mathrm{P}_\theta(\{y_1, \ldots, y_m\}) &= \mathbb{P}_\theta[Y_1 = y_1, \ldots, Y_m = y_m] \\ &= \theta^m(1-\theta)^{\sum_{i=1}^n y_i} = \exp\left(m\log(\theta) + \sum_{i=1}^n y_i \log(1-\theta)\right).\end{aligned}$$

Cette loi admet un rapport de vraisemblance monotone par rapport à la statistique $T(Y_1, \ldots, Y_n) \stackrel{\text{def}}{=} \sum_{i=1}^n Y_i$. Comme $\theta \mapsto \log(1-\theta)$ est une fonction décroissante de θ, le rapport de vraisemblance $p_{\theta'}/p_\theta(\mathbf{y})$ est une fonction croissante de $T(\mathbf{y})$ lorsque $\theta' < \theta$, donc le test U.P.P. de l'hypothèse $\theta \geq \theta_0$ consiste à rejeter cette hypothèse si T est trop grand. Ce test est d'ailleurs très intuitif : le nombre de tirage à effectuer avant de trouver m objets défectueux sera d'autant plus grand que la probabilité θ est petite. La statistique de test $T(\mathbf{Y})$, qui correspond ici au nombre de tirages à effectuer au delà de m pour obtenir m objets défectueux est distribué suivant une loi négative binomiale :

$$\mathrm{P}_\theta(T = t) = \binom{m+t-1}{m-1}\theta^m(1-\theta)^t.$$

Exemple 5.9 (Variance d'une loi gaussienne – suite de l'exemple 5.6)**:**
Soit $X = (X_1, \ldots, X_n)$ un n-échantillon gaussien $\mathcal{N}(0, \theta)$. Considérons l'hypothèse de base $H_0 = \{\theta \geq \theta_0\}$ et l'hypothèse alternative $H_1 = \{\theta < \theta_0\}$. On a vu précédemment que le rapport de vraisemblance $Z_{\theta,\theta'} = p_{\theta'}/p_\theta$ s'écrit en fonction de la statistique $T(X) = \sum_{i=1}^n X_i^2$,

$$Z_{\theta,\theta'}(\mathbf{x}) = \sqrt{\frac{\theta^2}{\theta'^2}} \exp\left\{-\frac{1}{2}(\theta'^{-2} - \theta^{-2})T(\mathbf{x})\right\}.$$

Pour $\theta' < \theta$, c'est une fonction strictement décroissante de $T(x)$ (ce qui correspond à l'intuition : si la variance empirique est faible, c'est vraisemblablement parce que la variance théorique est faible, et on rejettera alors H_0). La région critique d'un test de rapport de vraisemblance de type 5.2.1 (en changeant le sens des inégalités) s'écrit en fonction de la statistique T ,

$$\mathcal{X}_1 = \Big\{x : T(x) < d(\theta, \theta', c) = \frac{\log(\theta'^2/\theta^2) + 2\log c}{\theta^{-2} - \theta'^{-2}}\Big\}.$$

Ainsi, le test U.P.P. rejette H_0 lorsque $T(X_1, \ldots, X_n) \leq d$, où d est solution de l'équation :

$$\mathrm{P}_{\theta_0}(T \leq d) = \alpha.$$

Comme $T(X_1, \ldots, X_n)/\theta_0 = \sum_{i=1}^n X_i^2/\theta_0$ suit une loi du χ^2 centré à n degrés de liberté (voir l'annexe A.12), la constante critique du test est $\theta_0 x_n(\alpha)$ où $x_n(\alpha)$ est le quantile d'ordre α de la distribution χ^2_n.

Exemple 5.10 (Loi de Poisson)**:**
Soient $X_1, \ldots, X_n$ n variables distribuées suivant une loi de Poisson de paramètre θ, $\theta > 0$. Notons $X = (X_1, \ldots, X_n)$. La densité de probabilité (par rapport à la mesure de comptage) est donnée par

$$p_\theta(x_1, \ldots, x_n) = \frac{1}{\prod_{i=1}^n x_i!} \mathrm{e}^{-n\theta} \theta^{T(\mathbf{x})} \quad \text{où } T(\mathbf{x}) = \sum_{i=1}^n x_i, \mathbf{x} \in \mathbb{N}^n .$$

Notons que $T(X)$ suit une loi de Poisson de paramètre $n\theta$. Le rapport de vraisemblance s'écrit

$$Z_{\theta,\theta'}(x) = \mathrm{e}^{n(\theta-\theta') + T(\mathbf{x})\log(\theta'/\theta)},$$

c'est une fonction strictement croissante de $T(\mathbf{x})$ lorsque $\theta' > \theta$. Pour $c \geq 0$, un test U.P.P. de niveau α de l'hypothèse de base $H_0 : \theta \leq \theta_0$ contre l'hypothèse alternative $H_1 : \theta > \theta_0$ est donnée par (5.18), avec

$$\alpha = \sum_{j=\lfloor c \rfloor + 1}^{\infty} \frac{\mathrm{e}^{n\theta_0}(n\theta_0)^j}{j!}.$$

Il est intéressant de noter que la construction ci-dessus ne s'étend pas directement au cas des hypothèses bilatérales. Considérons $X_1, \ldots, X_n$ un n-échantillon i.i.d. d'une famille exponentielle associée à (h, T) de densité (par rapport à une mesure de domination μ)

$$p(x; \theta) = h(x) \exp(\phi(\theta)T(x) - \psi(\theta)),$$

où $\theta \to \phi(\theta)$ est une fonction croissante de θ. Supposons que $\mathrm{P}_\theta\left(\sum_{i=1}^n T(X_i) = c\right) = 0$ pour tout $\theta \in \Theta$ et pour tout c. En vertu du théorème de Neyman-Pearson, le test U.P.P. pour l'hypothèse de base $H_0 = \{\theta = \theta_0\}$ contre l'hypothèse $H_1 = \{\theta = \theta_1\}$ est non randomisé et admettra les régions critiques $T(x_1, \ldots, x_n) = \sum_{i=1}^n T(x_i) \geq c$ si $\theta_1 > \theta_0$ et $T(x_1, \ldots, x_n) \leq c$ si $\theta_0 < \theta_1$. On voit que la structure des tests U.P.P. est différente suivant que l'on considère des alternatives $\theta_1 > \theta_0$ et $\theta_1 < \theta_0$. C'est pourquoi il n'existe pas de test U.P.P. dans ce cadre.

5.6 Approche bayésienne

Il est aussi possible de considérer le problème de test dans un contexte bayésien. Cette approche consiste à prendre en compte la connaissance a priori sur le paramètre θ pour définir le risque. Plus précisément, contrairement à l'approche de Neyman-Pearson où l'on a cherché des procédures uniformément optimales sous contrainte (tests U.P.P.), on va chercher une procédure de décision optimale pour un risque intégré (voir le paragraphe 4.6). En d'autres termes, on cherche à exhiber un test qui soit une procédure de Bayes, au sens du paragraphe 4.6.

Soit $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$ un modèle statistique sur $\mathcal{X}$. On se donne π un *prior* sur $(\Theta, \mathcal{B}(\Theta))$: une mesure de probabilité représentant notre connaissance a priori. Soit maintenant δ une procédure de test pour l'hypothèse $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1$ c'est-à-dire, comme aux paragraphes précédents, une fonction de $\mathcal{X}$ dans $\{0, 1\}$. En utilisant comme d'habitude la fonction de perte $0-1$, le risque s'écrit toujours

$$R(\theta, \delta) = \mathbb{1}_{\Theta_0}(\theta)\mathbb{P}_{\theta_0}\Big[\delta(X) = 1\Big] + \mathbb{1}_{\Theta_1}(\theta)\mathbb{P}_{\theta_1}\Big[\delta(X) = 0\Big]$$

Le *risque intégré* de la procédure, telle quelle a été définie au paragraphe 4.6, s'écrit alors

$$r(\delta) = \int_\Theta R(\theta, \delta)\pi(d\theta) = \int_{\Theta_0} \mathbb{P}_\theta(\delta(X) = 1)\pi(d\theta) + \int_{\Theta_1} \mathbb{P}_\theta(\delta(X) = 0)\pi(d\theta).$$

Le test δ_π sera dit bayésien si, pour toute procédure (randomisée) de test δ, $r(\delta_\pi) \leq r(\delta)$. Comme suggéré par la notation, ce test optimal dépend du choix du prior π. Contrairement aux tests les plus puissants du cas non-bayésien, il est toujours possible de construire un test bayésien.

Supposons que le modèle statistique est dominé, $\mathrm{P}_\theta(dx) = p_\theta(x)\nu(dx)$, pour tout $\theta \in \Theta$, où ν est une mesure de référence sur $\mathcal{X}$. Rappelons que la loi a posteriori (voir les définitions 4.4.2 et 4.4.7) est une famille de probabilités $(\pi(\mathrm{d}\theta|x))_{x\in\mathcal{X}}$ indexées par l'observation $X = x$, telle que pour toute fonction mesurable bornée $\varphi : \Theta \times \mathcal{X} \to \mathbb{R}$,

$$\mathbb{E}(\varphi(\boldsymbol{\theta}, X)) \overset{\text{def}}{=} \int_\Theta \int_\mathcal{X} \varphi(\theta, x)p_\theta(x)\nu(\mathrm{d}x)\pi(\mathrm{d}\theta) = \int_\mathcal{X} \int_\Theta \varphi(\theta, x)\pi(\mathrm{d}\theta|x)m(x)\nu(\mathrm{d}x)$$

où $m(x)$ est la densité marginale de X sous le prior π, $m(x) = \int_\Theta p_\theta(x)\pi(\mathrm{d}\theta)$.

Considérons les probabilités a posteriori $\pi(\Theta_i|x)$ de chaque région $\Theta_i, i = \{0, 1\}$, sachant $X = x$,

$$\pi(\Theta_i|x) = \int_{\Theta_i} \pi(\mathrm{d}\theta|x), \quad i = 0, 1.$$

Considérons la règle de test consistant à choisir l'hypothèse « la plus probable a posteriori » :

$$\delta_\pi(x) = \begin{cases} 1 & \text{si } \pi(\Theta_1|x) > \pi(\Theta_0|x), \\ 0 & \text{sinon} \end{cases}$$

Dans la suite de ce paragraphe, on va montrer que δ_π est une procédure de Bayes pour le test de H_0 contre H_1.

Pour cela, on utilise la notion fondamentale de *risque a posteriori*. Elle fait intervenir l'espérance a posteriori (c'est-à-dire l'espérance conditionnelle, voir les définitions 4.4.10 et

4.4.12). Informellement, le risque a posteriori de l'action a est l'espérance de la perte encourue dans le futur en entreprenant a, sachant qu'on a observé $X = x$, ce qui nous permet d'utiliser la loi a posteriori $\pi(\mathrm{d}\theta|x)$ à la place du prior $\pi(\mathrm{d}\theta)$ pour calculer l'intégrale.

Définition 5.6.1 (risque a posteriori)**.** *Soit $(\mathcal{P} = \{P_\theta, \theta \in \Theta \subset \mathbb{R}^d\}, \pi)$ un modèle bayésien. On considère un problème de décision défini par un espace des actions $\mathcal{A}$, et une fonction de perte $L(\theta, a), \theta \in \Theta, a \in \mathcal{A}$. Soit $a \in \mathcal{A}$ une action.*

Le risque a posteriori *de l'action a pour le prior π, sachant l'observation x, que l'on notera $\rho_\pi(a, x)$, est l'espérance a posteriori de la perte $L(\boldsymbol{\theta}, a)$ (vue comme une fonction de $\boldsymbol{\theta}$), c'est-à-dire l'espérance conditionnelle de la perte $L(\boldsymbol{\theta}, a)$, sachant $X = x$,*

$$\rho_\pi(a, x) = \mathbb{E}(L(\boldsymbol{\theta}, a)|X = x) = \int_\Theta L(\theta, a)\pi(\mathrm{d}\theta|x).$$

L'intérêt de cette notion est de permettre une ré-écriture utile du risque intégré de toute procédure de décision δ, sous la forme d'une intégrale du risque a posteriori. En effet,

$$\begin{aligned} r(\delta) &\overset{\text{def}}{=} \int_\Theta R(\theta, \delta)\pi(\mathrm{d}\theta) \\ &= \int_\Theta \int_\mathcal{X} L(\theta, \delta(x))p_\theta(x)\nu(\mathrm{d}x)\pi(\mathrm{d}\theta) \\ &= \int_\mathcal{X} \underbrace{\int_\Theta L(\theta, \delta(x))\pi(\mathrm{d}\theta|x)}_{\rho_\pi(\delta(x),x)} m(x)\nu(\mathrm{d}x) \\ &= \int_\mathcal{X} \rho_\pi(\delta(x), x)m(x)\nu(\mathrm{d}x), \end{aligned} \tag{5.21}$$

Le lemme suivant prouve que, comme le suggère l'intuition, on a intérêt, étant donné une observation x, à prendre la décision $\delta(x)$ qui minimise l'espérance a posteriori de la perte.

Lemme 5.6.2 (Optimalité des décisions minimisant le risque a posteriori)
Soit $\delta^\star$ une procédure de décision à valeurs dans $\mathcal{A} \subset \mathbb{R}$, telle que pour tout $x \in \mathcal{X}$, $\delta^\star(x)$ minimise le risque a posteriori $\rho_\pi(\,\cdot\,|x)$, c'est-à-dire, telle que

$$\forall a \in \mathcal{A}, \quad \rho_\pi(a, x) \geq \rho_\pi(\delta^\star(x), x). \tag{5.22}$$

Alors la procédure $\delta^\star$ est une procédure de Bayes pour le prior π.

Démonstration. Soit δ une autre procédure de décision. Alors d'après(5.21), et en utilisant (5.22),

$$\begin{aligned} r(\delta) - r(\delta^\star) &= \int_\mathcal{X} \underbrace{\rho_\pi(\delta(x), x) - \rho_\pi(\delta^\star(x), x)}_{\geq 0} m(x)\nu(\mathrm{d}x) \\ &\geq 0. \end{aligned}$$

■

Dans le cas des tests d'hypothèses, l'espace des actions est $\mathcal{A} = \{0, 1\}$, ce qui simplifie le problème de minimisation intervenant dans le lemme 5.6.2 (on cherche un minimiseur dans l'ensemble $\{0, 1\}$).

Proposition 5.6.3
Le test δ_π donné par

$$\delta_\pi(x) = \begin{cases} 1 & si\ \pi(\Theta_1|x) > \pi(\Theta_0|x), \\ 0 & sinon \end{cases} \tag{5.23}$$

est un test de Bayes pour le test de l'hypothèse H_0 contre H_1.

Démonstration. Dans le cadre des tests, la fonction de coût est, pour $a \in \{0, 1\}$,

$$L(\theta, a) = \begin{cases} 0 & \text{si } \theta \in \Theta_a \\ 1 & \text{sinon} \end{cases}.$$

Ainsi, le risque a posteriori $\rho_\pi(a, x)$ s'écrit, pour $a \in \{0, 1\}$,

$$\begin{aligned} \rho_\pi(a, x) &= \int_{\Theta_0} L(\theta, a)\pi(\mathrm{d}\theta|x) + \int_{\Theta_1} L(\theta, a)\pi(\mathrm{d}\theta|x) \\ &= \begin{cases} \int_{\Theta_0} 1\ \pi(\mathrm{d}\theta|x) + \int_{\Theta_1} 0\ \pi(\mathrm{d}\theta|x) & \text{si } a = 1 \\ \int_{\Theta_0} 0\ \pi(\mathrm{d}\theta|x) + \int_{\Theta_1} 1\ \pi(\mathrm{d}\theta|x) & \text{si } a = 0 \end{cases} \\ &= \begin{cases} \pi(\Theta_0|x) & \text{si } a = 1 \\ \pi(\Theta_1|x) & \text{si } a = 0 \end{cases} \end{aligned}$$

Par conséquent, on a

$$\rho_\pi(1, x) < \rho_\pi(0, x) \iff \pi(\Theta_0|x) < \pi(\Theta_1|x).$$

Ainsi, la règle de décision définie par (5.23) satisfait (5.22) et le lemme 5.6.2 permet de conclure.

■

Passons à la détermination pratique du test de Bayes (5.23) : supposons que la loi a priori admet une densité $\pi(\theta)$ par rapport à une mesure de référence μ sur $(\Theta, \mathcal{B}(\Theta))$, $\pi(d\theta) = \pi(\theta)\mu(d\theta)$. Alors la densité marginale de X par rapport à la mesure de référence ν sur $\mathcal{X}$ s'écrit $m(x) = \int_\Theta p_\theta(x)\pi(\theta)\mu(d\theta)$. De plus, en notant $\pi(\theta|x)$ la densité de la loi a posteriori du paramètre par rapport à μ, on a

$$\pi(\theta|x) = \begin{cases} p_\theta(x)\pi(\theta)/m(x) & \text{si } m(x) \neq 0, \\ 0 & \text{sinon} \end{cases}$$

Le test bayésien prend alors la forme

$$\delta_\pi(x) = \begin{cases} 1 & \text{si } \int_{\Theta_1} p_\theta(x)\pi(\theta)\mu(d\theta) > \int_{\Theta_0} p_\theta(x)\pi(\theta)\mu(d\theta), \\ 0 & \text{sinon.} \end{cases}$$

Notons que les intégrales ci-dessus définissant le test ne sont autres que la vraisemblance $p_\theta(x)$, intégrée sur la région considérée (Θ_0 ou Θ_1), sous la loi a priori π. On choisit donc l'alternative qui a la plus grande « vraisemblance intégrée » (sous le prior π).

Exemple 5.11 (Paramètre d'une loi binomiale)**:**
Soit X une v.a. distribuée selon une variable binomiale $\mathcal{B}(n,\theta)$ où $\theta \in (0,1)$, pour $x \in \{0,\dots,n\}$,

$$\mathrm{P}_\theta(X=x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

Prenons pour loi a priori la loi uniforme sur $[0,1]$. Posons $\Theta_0 = [0,1/2]$. La probabilité a posteriori de l'événement $\{\boldsymbol{\theta} \in \Theta_0\}$ est donné par :

$$\begin{aligned}\pi(\Theta_0|x) &= \frac{\int_0^{1/2}\binom{n}{x}\theta^x(1-\theta)^{n-x}\mathrm{d}\theta}{\int_0^1\binom{n}{x}\theta^x(1-\theta)^{n-x}\mathrm{d}\theta}\\ &= \frac{\int_0^{1/2}\theta^x(1-\theta)^{n-x}\mathrm{d}\theta}{B(x+1,n-x+1)}\\ &= \frac{(1/2)^{n+1}}{B(x+1,n-x+1)}\sum_{k=0}^{n-x}\frac{(n-x)!\,x!}{(n-x-k)!\,(x+k+1)!},\end{aligned}$$

où $B(a,b) = \int_0^1\theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta$ est la fonction Bêta (tabulée dans les librairies numériques de math). On montre facilement la dernière égalité par récurrence en effectuant une intégration par partie. La dernière expression donc calculable numériquement. Cette procédure de test a été proposée par Laplace pour tester l'hypothèse qu'à la naissance, le nombre d'enfants de sexe masculin excédait le nombre d'enfants du sexe opposé.

Exemple 5.12 (Moyenne d'une loi gaussienne)**:**
Supposons que l'observation $X = (X_1,\dots,X_n)$ est distribué suivant une loi gaussienne de moyenne θ et de variance σ^2.

$$p(x_1,\dots,x_n|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}^n}\exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^n(x_i-\theta)^2\right).$$

Prenons comme loi a priori π une loi gaussienne de moyenne μ et de variance τ^2. Cette loi a priori est conjuguée et la loi a posteriori est une loi gaussienne de moyenne $\mu(\bar{x})$ et de variance ω^2,

$$\mu(\bar{x}) = \frac{\sigma^2\mu/n + \tau^2\bar{x}}{\sigma^2/n+\tau^2} \quad \text{et} \quad \omega^2 = \frac{\sigma^2\tau^2/n}{\sigma^2/n+\tau^2}$$

où $\bar{x} = n^{-1}\sum_{i=1}^n x_i$. Pour tester l'hypothèse $H_0 : \theta < 0$, nous calculons la probabilité a posteriori :

$$\mathbb{P}(\boldsymbol{\theta}<0|x) = \mathbb{P}\left(\frac{\boldsymbol{\theta}-\mu(\bar{X})}{\omega} < -\mu(\bar{X})/\omega \;\middle|\; x\right) = \Phi(-\mu(\bar{X})/\omega),$$

où Φ est la fonction de répartition de la loi gaussienne standard.

5.7 Lien entre approche bayésienne et approche de Neyman-Pearson

Dans un cadre d'un test d'hypothèses simples, on peut considérer que l'espace des paramètres est réduit à l'ensemble à deux éléments $\{\theta_0,\theta_1\}$. Spécifier la loi a priori pour un test

d'hypothèse simple revient simplement à choisir une probabilité a priori pour l'hypothèse de base $\{\theta = \theta_0\}$. On note $\pi_0 = \pi(\{\theta_0\})$. Le test de bayes s'écrit donc

$$\delta_\pi(x) = \begin{cases} 1 & \text{si } \pi(\theta_1|x) > \pi(\theta_0|x), \\ 0 & \text{sinon} \end{cases} \tag{5.24}$$

De plus, a loi a posteriori s'écrit alors :

$$\mathbb{P}(\boldsymbol{\theta} = \theta_0|x) = \pi(\theta_0|x) = \frac{\pi_0 p(x|\theta_0)}{\pi_0 p(x|\theta_0) + (1-\pi_0)p(x|\theta_1)},$$

et nous avons :

$$\pi(\theta_1|x) > \pi(\theta_0|x) \quad \Leftrightarrow \quad Z(x) = \frac{p_1(x)}{p_0(x)} > \frac{\pi_0}{1-\pi_0}.$$

Le test bayésien consiste donc à choisir l'hypothèse 1 si le *rapport de vraisemblance* $Z(x) = p_1(x)/p_0(x)$ excède un seuil dont la valeur dépend de la probabilité a priori π de l'hypothèse de base. Autrement dit,

$$\delta_\pi(x) = \begin{cases} 1 & \text{si } Z(x) > \pi_0/(1-\pi_0), \\ 0 & \text{sinon} \end{cases} \tag{5.25}$$

Ceci suggère un lien avec le cadre des tests U.P.P., qui nous explicitons ci-dessous.

Proposition 5.7.1
Soit $\mathcal{P} = \{\mathrm{P}_{\theta_0}, \mathrm{P}_{\theta_1}\}$ et considérons le problème de tests d'hypothèses simples $H_0 : \{\theta = \theta_0\}$ contre $H_1 : \{\theta = \theta_1\}$. Soit $\pi_0 \in (0,1)$ la probabilité a priori de $\{\theta_0\}$, de sorte que la loi a priori est la probabilité discrète $(\pi_0, 1-\pi_0)$. On considère le test de Bayes δ_π associé, défini par (5.24). Soit $\alpha = R(\theta_0, \delta_\pi) = \mathbb{P}_{\theta_0}(\delta_\pi(X) = 1)$ son risque de première espèce.

Alors, δ_π est U.P.P. de niveau α.

Démonstration.
Première preuve. On a montré précédemment que δ_π est défini de manière équivalente par (5.25). C'est donc un test de rapport de vraisemblance de type (5.6), avec $c = \pi_0/(1-\pi_0)$. Comme on a supposé que le test est de niveau α, le théorème 5.2.1 s'applique et δ_π est U.P.P. de niveau α.
Deuxième preuve (directe) : Nous avons, pour tout test δ,

$$r(\delta_\pi) = \pi\alpha + (1-\pi)R(\theta_1, \delta_\pi) \leq r(\delta) = \pi R(\theta_0, \delta) + (1-\pi)R(\theta_1, \delta),$$

et par conséquent :

$$0 \leq \pi(\alpha - R(\theta_0, \delta)) \leq (1-\pi)(R(\theta_1, \delta) - R(\theta_1, \phi_\pi)).$$

■

Exemple 5.13 (Classification binaire : discrimination linéaire)**:**
Considérons un test d'hypothèse simple, où les lois $p_0(x)$ et $p_1(x)$ sont des lois gaussiennes multidimensionnelles de paramètres $\theta_0 = (m_0, \Sigma_0)$ et $\theta_1 = (m_1, \Sigma_1)$, où m_i, $i = 0, 1$ sont les moyennes et Σ_i, $i = 0, 1$ sont les matrices de covariance, supposées ici non singulières :

$$p_i(x) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma_i)}} \exp\left(-\frac{1}{2}(x-m_i)^T \Sigma_i^{-1}(x-m_i)\right).$$

En notant π la probabilité a priori de $\boldsymbol{\theta} = \theta_0$, la loi de l'observation X est un mélange de gaussiennes de proportion π, $f_\pi(x) = \pi p_0(x) + (1-\pi)p_1(x)$ et la règle de Bayes est alors de la forme :

$$\phi_\pi(x) = \begin{cases} 1 & \text{si } (1-\pi)p_1(x) > \pi p_0(x), \\ 0 & \text{sinon} \end{cases}$$

En prenant les logarithmes, on remarque que $\phi_\pi(x) = 1$ si et seulement si :

$$(x-m_1)^T\Sigma_1^{-1}(x-m_1) - 2\log(1-\pi) + \log(\det(\Sigma_1)) \\ < (x-m_0)^T\Sigma_0^{-1}(x-m_0) - 2\log(\pi) + \log(\det(\Sigma_0)).$$

$r_i^2 = (x-m_i)^T\Sigma_i^{-1}(x-m_i)$ est le carré de *la distance de Mahalanobis* entre x et m_i dans la classe i, une distance couramment utilisée en reconnaissance des formes. En fonction de cette distance, la règle de Bayes est donc de la forme :

$$\phi_\pi(x) = \begin{cases} 1 & \text{si} \quad r_1^2 < r_0^2 - 2\log(\pi/(1-\pi)) + \log(\det(\Sigma_0)/\det(\Sigma_1)), \\ 0 & \text{sinon} \end{cases}.$$

Lorsque $\pi = 1/2$ et que $\Sigma_0 = \Sigma_1 = \Sigma$, la règle devient simplement :

$$\phi_\pi(x) = \begin{cases} 1 & \text{si} \quad r_1^2 < r_0^2, \\ 0 & \text{sinon} \end{cases}$$

et on choisit donc la « classe » i dont la distance de Mahalanobis de x à m_i dans la classe i est la plus petite. Lorsque $\Sigma_1 = \Sigma_0 = \Sigma$, on montre facilement que la règle de Bayes est équivalente à une règle de discrimination linéaire :

$$\phi_\pi(x) = \begin{cases} 1 & \text{si} \quad a^Tx + a_0 > 0, \\ 0 & \text{sinon} \end{cases} \tag{5.26}$$

où $a = \Sigma^{-1}(m_1 - m_0)$, $a_0 = 2\log((1-\pi)/\pi) + m_0^T\Sigma^{-1}m_0 - m_1^T\Sigma^{-1}m_1$. Nous avons visualisé dans les figures 5.4 et 5.5 deux échantillons de 500 variables gaussiennes indépendantes bi-dimensionnelles de moyennes $\mu_0 = [1.5, 0]$ et $\mu_1 = [0, -10]$ (figure 5.4), $\mu_0 = [1.5, 0]$ et $\mu_1 = [0, -2]$ (figure 5.5) :

$$\Sigma = \begin{bmatrix} 4 & 1 \\ 1 & 1 \end{bmatrix}. \tag{5.27}$$

Dans le premier cas, les deux classes sont clairement séparées à l'inverse du second cas, où la distinction des classes est plus difficile à faire. Dans les deux cas, le test bayésien consiste à calculer la droite (5.26), et à accepter H_0 où H_1 suivant que l'observation x se trouve dans l'un ou l'autre des deux demi-plans délimités par cette droite de séparation.

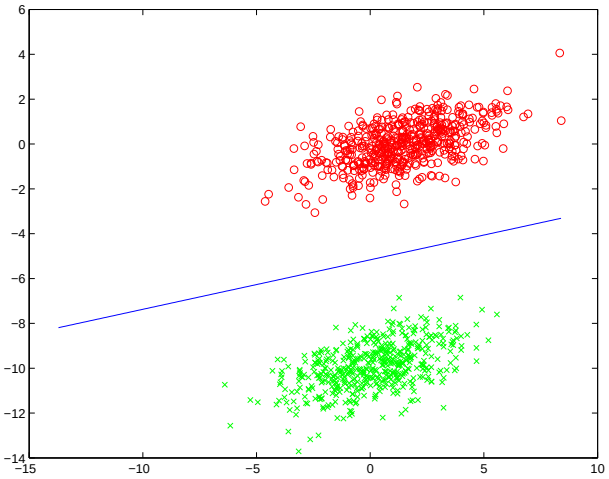


FIGURE 5.4 – Échantillons de loi gaussienne bi-dimensionnelles de moyenne $\mu_0 = [1.5, 0]$ et $\mu_1 = [0, -10]$ et de matrice de covariance (5.27). Le test consiste à choisir H_0 ou H_1 suivant que l'observation se trouve au-dessous ou au dessus de la droite de séparation

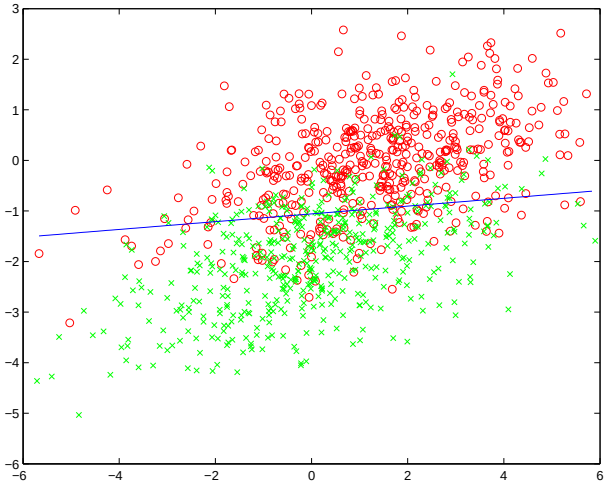


FIGURE 5.5 – Échantillons de loi gaussienne bi-dimensionnelles de moyenne $\mu_0 = [1.5, 0]$ et $\mu_1 = [0, -2]$ et de matrice de covariance (5.27). Le test consiste à choisir H_0 ou H_1 suivant que l'observation se trouve au-dessous ou au-dessus de la droite de séparation

Chapitre 6

Intervalles et régions de confiance

Dans toute la suite de ce chapitre, on se donne un modèle statistique $\{\mathrm{P}_\theta, \theta \in \Theta\}$ et on observe $X \sim \mathrm{P}_\theta$ avec θ inconnu. On note $\mathcal{X}$ l'espace des observations.

6.1 Régions et intervalles de confiance

Déterminer une région de confiance pour le paramètre inconnu θ d'une loi P_θ est généralement la deuxième étape d'une analyse de données : on demande d'abord un estimateur $\widehat{\theta}$, puis on se demande quelle confiance accorder à cette estimation, et on aimerait un intervalle de confiance autour de $\widehat{\theta}$. On demande donc de fournir une région de Θ qui contienne le vrai paramètre θ avec une grande probabilité. On n'aura pas besoin d'hypothèses particulières sur Θ pour les résultats qui suivent, mais supposons pour commencer que $\Theta = \mathbb{R}$ pour fixer les idées. Bien sûr, si l'on fournit un intervalle I fixé, l'affirmation « $\mathbb{P}(\theta \in I) \geq 0.95$ » n'a pas de sens dans le cadre classique où θ n'est pas une variable aléatoire mais un nombre. Pourtant, si l'on considère que l'intervalle I est construit en fonction des données, $I = I(X)$, alors les bornes de l'intervalle $m(X), M(X)$ sont des variables aléatoires et on peut écrire

$$\theta \in I(X) \Leftrightarrow \{m(X) \leq \theta, M(X) \geq \theta\},$$

qui représente bien un événement au sens probabiliste. Dans un cadre plus général (Θ un ensemble quelconque), on va construire une *région de confiance* $\delta(X) \subset \Theta$, fonction des données observées.

Souvent, la quantité d'intérêt n'est pas θ lui-même mais une fonction de θ, $g(\theta) \in \mathbb{R}$. Par exemple, dans le cas gaussien, $\theta = (\mu, \sigma^2)$ et on peut vouloir simplement un intervalle de confiance concernant μ, de sorte que l'on posera $g(\theta) = \mu$.

Définition 6.1.1 (Intervalle et région de confiance)**.** *On se donne $\alpha \in]0,1[$. La quantité $1-\alpha$ est appelée* niveau de confiance.

(1) *Une région de confiance au niveau $1-\alpha$ pour le paramètre θ est une région aléatoire $\delta(X) \subset \Theta$, telle que*

$$\inf_{\theta \in \Theta} \mathbb{P}_\theta \left(\theta \in \delta(X)\right) = 1 - \alpha, \tag{6.1}$$

(2) *Soit $g : \Theta \to \mathbb{R}$ une fonction à valeurs réelles. Un intervalle de confiance de niveau $1-\alpha$ pour le paramètre $g(\theta)$ est un intervalle $I(X) = [m(X), M(X)]$ d'extrémités aléatoires vérifiant*

$$\inf_{\theta \in \Theta} \mathbb{P}_\theta \left(g(\theta) \in [m(X), M(X)]\right) = 1 - \alpha, \tag{6.2}$$

Il est courant d'exprimer $1-\alpha$ sous la forme d'un pourcentage. Si la niveau de confiance $1-\alpha$ est de 0.95 alors nous dirons que l'intervalle $[m(X), M(X)]$ est intervalle de confiance à 95%.

Remarque 6.1.2. *Un intervalle de confiance définit un cas particulier de région de confiance. En effet, dire que $I(X)$ est un intervalle de confiance de niveau $1-\alpha$ pour $g(\theta)$ revient à dire que la région*

$$\delta(X) = \{\theta : g(\theta) \in I(X)\} = g^{-1}(I(X))$$

est une région de confiance pour θ de niveau $1-\alpha$.

Ce formalisme mathématique a une interprétation simple. On retiendra :

> **Une région de confiance de niveau $1-\alpha$ est une région déterminée en fonction des données, telle que, quellle que soit la loi P_θ des observations (dans les limites du modèle statistique considéré), la région contienne le paramètre θ avec probabilité $1-\alpha$.**

Il est important d'insister sur le fait *absolument essentiel* que les conditions (6.1) (*resp.* (6.2)) impliquent que l'inégalité

$$\begin{aligned} &\mathbb{P}_\theta\left(\theta \in \delta(X)\right) \geq 1-\alpha \\ \textit{(resp.} \quad &\mathbb{P}_\theta\left(\theta \in [m(X), M(X)]\right) \geq 1-\alpha \quad \textit{)} \end{aligned}$$

doit être vérifiée pour *toutes les valeurs possibles du paramètre θ*. Cette contrainte peut sembler assez forte, mais nous allons voir dans la suite qu'il est possible de la satisfaire dans de nombreux cas d'intérêt pratique. En réalité, cette définition n'est pas très restrictive : par exemple, il n'est pas difficile de définir une région de confiance de niveau de confiance arbitrairement grand. Par exemple, $\delta = \Theta$ est de niveau de confiance 1 mais ne présente aucun intérêt. Ce qui fait la valeur pratique d'une région de confiance de niveau de confiance donné est sa *taille*. La façon la plus naturelle pour comparer la taille de deux régions est la relation d'inclusion : δ_1 est plus petite que δ_2 si $\delta_1 \subseteq \delta_2$. Cette relation étant partielle, elle n'est pas satisfaisante pour définir une région de confiance optimale.

Dans les situations les plus sympathiques, il est possible de choisir les statistiques $m(X)$ et $M(X)$ de telle sorte que $\mathbb{P}_\theta\left(g(\theta) \in [m(X), M(X)]\right)$ soit en fait indépendante de θ. Cela sera le cas (cf. les exemples 6.1 et 6.2) lorsque nous chercherons à déterminer les intervalles de confiance pour la moyenne d'un échantillon gaussien, que la variance soit connue ou inconnue. Dans les cas plus complexes, il n'est pas possible de calculer *exactement* la quantité $\mathbb{P}_\theta\left(g(\theta) \in [m(X), M(X)]\right)$, ni de trouver une borne inférieure à cette quantité. Dans ce cas-là (fréquent, mais qui sort du cadre de ce cours), les méthodes *asymptotiques* fournissent des intervalles de confiance valides dans la limite des grands échantillons.

6.2 Lien avec la théorie de la décision

Le problème de la détermination d'une région de confiance est un problème de décision où l'espace des actions $\mathcal{A}$ est l'ensemble $\mathcal{P}(\Theta)$ des parties de Θ. Comme dans le cas des tests, la fonction de coût considérée sera la perte 0-1, où l'on perd 1 si l'on « se trompe » et 0 sinon, c'est-à-dire, pour $I \subset \Theta$,

$$L(\theta, I) = \begin{cases} 1 & \text{si } \theta \notin I \\ 0 & \text{si } \theta \in I \end{cases}$$

Notons comme dans les chapitres précédents $\delta : \mathcal{X} \to \mathcal{A} = \mathcal{P}(\Theta)$ la procédure de décision permettant de déterminer une région de confiance. Comme dans le cas des tests statistiques, la fonction de perte utilisée vaut 0 ou 1, suivant que le paramètre à localiser appartient ou non à la région de confiance $\delta(X)$:

$$R(\theta, \delta) = \mathbb{P}_\theta(\theta \notin \delta(X)) = 1 - \mathbb{P}_\theta(\theta \in \delta(X)) \ .$$

Cette dernière expression du risque est écrite en fonction de $\mathbb{P}_\theta(\theta \in \delta(X))$, qui s'appelle la *probabilité de couverture*. La moins bonne probabilité de couverture associée à une région de confiance est, d'après la définition 6.1.1, son *niveau de confiance* : $\inf_{\theta\in\Theta} \mathbb{P}_\theta(\theta \in \delta(X)) = 1-\alpha$. Autrement dit, une région de confiance est de niveau $1-\alpha$ si son risque « maximum » vaut α,

$$\sup_{\theta\in\Theta} R(\theta, \delta) = \alpha.$$

Ceci suggère un lien entre la construction des régions de confiance et la construction de tests statistiques de niveau α vus au chapitre 5 (voir en particulier la définition 5.1.1). Nous détaillerons les liens existant entre tests et intervalles de confiance au paragraphe 6.4.

Ces définitions ne nous disent pas comment construire une région de confiance en pratique. Ceci n'est pas toujours possible mais certaines situations permettent de le faire. C'est l'objet du paragraphe suivant.

6.3 Construction à l'aide de fonctions pivotales

Les fonctions pivotales sont l'outil de base pour la construction d'intervalles de confiance. Commençons par un exemple simple

Exemple 6.1 (Intervalle de confiance pour un échantillon gaussien de variance connue)**:**
Soit $X = (X_1, \ldots, X_n)$ un n-échantillon i.i.d. d'une loi gaussienne $\mathcal{N}(\mu, \sigma^2)$ de *variance connue*. Nous cherchons à construire un intervalle de confiance $[m(X), M(X)]$ pour la moyenne μ de niveau de confiance $1-\alpha$, c'est-à-dire tel que

$$\mathbb{P}_\mu\left(\mu \in [m(X), M(X)]\right) = 1 - \alpha. \tag{6.3}$$

Dans cet exemple élémentaire, il est aisé de construire un tel intervalle : remarquons en effet que la variable aléatoire $Z = \sqrt{n}(\bar{X}_n - \mu)/\sigma$, où $\bar{X}_n$ est la moyenne empirique, est distribué suivant une loi gaussienne centrée réduite : la distribution de Z est indépendante de la valeur du paramètre μ. Soit Φ la fonction de répartition d'une loi gaussienne centrée réduite :

$$\Phi(z) = \int_{-\infty}^{z} (2\pi)^{-1/2} \exp(-x^2/2) dx \ .$$

Notons, pour $0 \leq \alpha \leq 1$, $z(\alpha)$ le α-quantile défini ici par

$$\Phi(z(\alpha)) = \alpha \ .$$

En particulier, pour les valeurs usuelles de α, nous avons $z(1-\alpha/2) = 1,96$ si $\alpha = 0,05$ et $z(1-\alpha/2) = 3$ si $\alpha = 0,01$. L'intervalle symétrique $[-z(1-\alpha/2), z(1-\alpha/2)]$ vérifie alors

$$\mathbb{P}_\mu\left(\sqrt{n}(\bar{X}_n - \mu)/\sigma \in [-z(1-\alpha/2), z(1-\alpha/2)]\right) = 1 - \alpha \ .$$

Autrement dit

$$\mathbb{P}_\mu \left(\mu \in [\bar{X}_n - \sigma\, z(1-\alpha/2)/\sqrt{n}, \bar{X}_n + \sigma\, z(1-\alpha/2)/\sqrt{n}] \right) = 1 - \alpha\ .$$

On vient de montrer que $m(X) = \bar{X}_n - \sigma z(1-\alpha/2)/\sqrt{n}$ et $M(X) = \bar{X}_n + \sigma z(1-\alpha/2)/\sqrt{n}$ sont les extrémités d'un intervalle de confiance de niveau de confiance $1-\alpha$ pour μ : (6.3) est valide pour tout $\mu \in \mathbb{R}$. On remarque au passage que l'intervalle de confiance bilatéral a pour diamètre $2\sigma z(1-\alpha/2)/\sqrt{n}$ qui tend vers 0 quand $n \to \infty$ à un niveau de confiance donné.

Dans cet exemple, on peut aussi construire une borne de confiance inférieure de niveau $1-\alpha$ donné : $\bar{X}_n - \sigma z(1-\alpha)/\sqrt{n}$ ou une borne de confiance supérieure de niveau $1-\alpha$ donné : $\bar{X}_n + \sigma z(1-\alpha)/\sqrt{n}$.

Dans l'exemple ci-dessus, on a trouvé une fonction $\varphi(X, \theta)$ (avec $\theta = \mu$), définie par

$$\varphi(X, \theta) = \sqrt{n}(\bar{X}_n - \theta)/\sigma,$$

telle que la loi de la variable aléatoire $Z = \varphi(X, \theta)$ ne dépende pas de θ (dans l'exemple, Z suit une loi normale standard, quel que soit θ). La fonction φ sera appelée *fonction pivotale*.

Définition 6.3.1 (Fonction Pivotale)**.** *On dit qu'une fonction $\varphi : \mathcal{X} \times \Theta \to \mathbb{R}$ est pivotale si quelle que soit la loi P_θ de l'observation X, la loi de $\varphi(X, \theta)$ ne dépend pas de θ.*

L'avantage d'utiliser une fonction pivotale est que l'on peut manipuler $Z = \varphi(X, \theta)$ dont la loi est constante, et généralement connue. Ainsi, sans connaître θ, on peut trouver $A \subset \mathbb{R}$ (un ensemble mesurable) tel que

$$1 - \alpha = \mathbb{P}\left(Z \in A\right) = \mathbb{P}_\theta\left(\varphi(X, \theta) \in A\right), \quad \text{pour tout} \quad \theta \in \Theta\ .$$

Il s'en suit que, pour tout ensemble A ainsi choisi, la région définie par

$$\delta(X) = \{\theta \ :\ \varphi(X, \theta) \in A\}$$

est une région de confiance de niveau de confiance $1-\alpha$.

Pour trouver en pratique un ensemble A, la notion de quantile, déjà évoqué dans l'exemple 6.1, sera utile en général.

Définition 6.3.2 (Quantile)**.** *Soit Z une v.a. réelle. Pour $p \in (0, 1)$, le nombre z est un quantile d'ordre p de la loi de Z si*

$$\mathbb{P}(Z \leq z) = p\ .$$

Si la loi de Z a des "trous" dans son support, par exemple si Z est une variable discrète, cette équation ne définit pas toujours z de manière unique et n'a pas toujours de solution. Pour simplifier la discussion, supposons l'existence et l'unicité du quantile d'ordre p de la variable $Z = \varphi(X, \theta)$, que l'on notera $z(p)$ – il suffira sinon de choisir le quantile d'ordre p le plus favorable (le plus petit ou le plus grand) en terme de taille de région de confiance obtenue – de telle sorte que l'on puisse définir une fonction $z \mapsto z(p)$, appelée *fonction quantile*. Pour déterminer un ensemble A qui convienne, on peut par exemple prendre $A = (-\infty, z(1-\alpha)]$ ou $A = (z(\alpha), \infty)$. Si la loi de $g(X, \theta)$ est symétrique, on a $z(1-p) = -z(p)$ et on peut aussi choisir $A = (-z(1-\alpha/2), z(1-\alpha/2)]$. En fait plus généralement, pour tout p_1 et p_2 tels que $0 \leq p_1 < p_2 \leq 1$ et $1 - \alpha = p_2 - p_1$, on peut choisir $A = (z(p_1), z(p_2)]$.

Le choix de A est guidé soit par un objectif particulier soit par la volonté de minimiser la taille de la région de confiance, dans un sens à préciser. On s'intéressera le plus souvent au cas de l'estimation par intervalle où la région de confiance recherchée s'écrit en fonction d'un paramètre scalaire $g : \Theta \to \mathbb{R}$:

$$\delta(X) = \{\theta \ : \ g(\theta) \in I(X)\} \ ,$$

où l'intervalle $I(X)$ est sous l'une des trois formes suivantes :

(i) $I(X) = [m(X), \infty)$: $m(X)$ est une *borne inférieure de confiance.*

(ii) $I(X) = (\infty, M(X)]$: $M(X)$ est une *borne supérieure de confiance*

(iii) $I(X) = [m(X), M(X)]$ est un *intervalle de confiance bilatéral.*

Pour déterminer des intervalles de confiance de niveau de confiance $(1-\alpha)$, on fera donc appel à des fonctions pivotales de la forme $(\theta, X) \mapsto \varphi(X, g(\theta))$.

Pourquoi prendre l'intervalle symétrique dans l'exemple 6.1 ? Il se trouve que dans le cas de cet exemple, ce choix était le bon, en vertu du lemme suivant.

Lemme 6.3.3
Soit $f : \mathbb{R} \to \mathbb{R}^+$ une densité unimodale*, c'est-à-dire n'admettant qu'un seul maximum, appelé mode de f. On suppose que le mode de f est nul et que f est croissante sur $\mathbb{R}_-$ et symétrique. Soit une v.a. X de densité f. Pour tout $\alpha > 0$, un couple (a, b) tel que $a = -b$ minimise la longueur $b - a$ des intervalles vérifiant $\mathbb{P}(X \in [a, b]) = 1 - \alpha$.*

Considérons maintenant un exemple un peu plus réaliste : on cherche toujours à estimer la moyenne d'un échantillon gaussien, mais on ne connaît pas la variance σ^2.

Exemple 6.2 (Intervalle de confiance pour la moyenne à variance inconnue)**:**
Soit $X = (X_1, \ldots, X_n)$ un n-échantillon i.i.d. de v.a. gaussiennes de moyenne $\mu \in \mathbb{R}$ et de variance $\sigma^2 > 0$ inconnue. On cherche un intervalle de confiance de niveau $1 - \alpha$ pour le paramètre μ.

Lorsque la variance est inconnue, nous allons « remplacer » σ^2 par son estimateur empirique non biaisé

$$\widehat{s}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2.$$

Considérons alors la variable aléatoire

$$T = \varphi(X, \mu) = \sqrt{n}(\bar{X}_n - \mu)/\widehat{s}_n.$$

Le théorème A.12.24 montre que la variable aléatoire $T = \varphi(X, \mu)$, quels que soient les paramètres μ et σ^2, est distribuée suivant une loi de Student à $(n-1)$ degrés de liberté, $T \sim \mathbf{T}(n-1)$. Notons $t_{n-1}(p)$ le p-quantile de la loi $\mathbf{T}(n-1)$. Remarquons que la loi de Student est unimodale et que son mode est 0 (ce qui justifie, comme dans le cas précédent, de considérer des intervalles symétriques). Par conséquent,

$$\mathbb{P}_{\mu,\sigma^2} \left(-t_{n-1}(1-\alpha/2) \leq \varphi(X, \mu) \leq t_{n-1}(1-\alpha/2) \right) = 1 - \alpha.$$

En résolvant la relation précédente par rapport à μ, nous obtenons, pour tout μ, σ^2 :

$$\mathbb{P}_{\mu,\sigma^2} \left(\mu \in [m, M] \right) = 1 - \alpha,$$

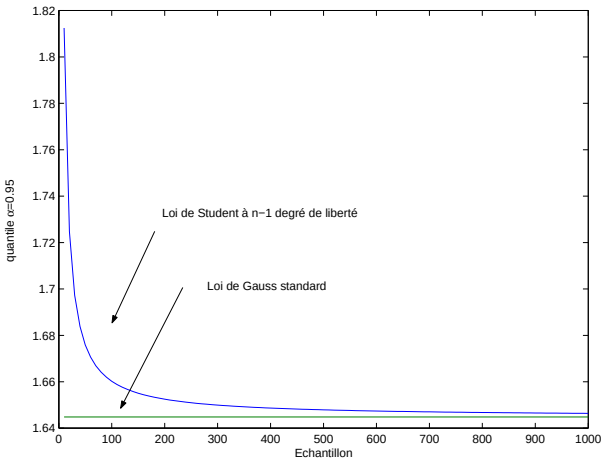


FIGURE 6.1 – Quantile d'ordre $p = 0.95$ pour une loi de Student à $(n-1)$ degrés de liberté et d'une loi gaussienne standard, en fonction de la taille n de l'échantillon.

avec

$$m(X) = \bar{X}_n - \widehat{s}_n t_{n-1}(1-\alpha/2)/\sqrt{n} \quad \text{et} \quad M(X) = \bar{X}_n + \widehat{s}_n t_{n-1}(1-\alpha/2)/\sqrt{n}.$$

Pour évaluer en pratique les quantiles de la loi de Student, on peut utiliser soit des tables ou (ce qui est plus pratique) des logiciels statistiques. Nous avons représenté dans la figure 6.1 le quantile d'ordre $p = 0.95$ des lois de Student à $(n-1)$ degrés de liberté et de la loi gaussienne centrée réduite. Nous voyons sur ce graphique que dès que la taille de l'échantillon dépasse $n \geq 100$, les valeurs $t_{n-1}(0.95)$ et $z(0.95)$ sont très proches.

Exemple 6.3 (Intervalle de confiance pour la variance)**:**
On considère encore une fois $X = (X_1, \ldots, X_n)$ un n-échantillon gaussien $\mathcal{N}(\mu, \sigma^2)$, où μ et σ^2 sont inconnus. On cherche cette fois un intervalle de confiance pour la variance σ^2. La variable aléatoire $V = \varphi(X, \sigma^2) = (n-1)\widehat{s}_n^2/\sigma^2$ est distribuée suivant une loi de χ^2 à $n-1$ degrés de liberté et peut être utilisée comme quantité pivotale. Si nous notons $x_{n-1}(p)$ le quantile d'ordre p de la loi χ^2_{n-1} et si nous prenons $\alpha_1 + \alpha_2 = \alpha$, alors, pour tout (μ, σ^2) :

$$\mathrm{P}_{\mu,\sigma^2}\Big(x_{n-1}(\alpha_1) \leq V(\cdot, \sigma^2) \leq x_{n-1}(1-\alpha_2)\Big) = 1 - \alpha.$$

En résolvant l'équation précédente par rapport à σ^2, nous obtenons donc que :

$$\Big[(n-1)\widehat{s}_n^2/x_{n-1}(1-\alpha_2), (n-1)\widehat{s}_n^2/x_{n-1}(\alpha_1)\Big]$$

est un intervalle de confiance pour σ^2 de niveau de confiance $1-\alpha$. La longueur de cet intervalle est aléatoire, $L_n(s^2, \alpha_1, \alpha_2)$.

Il est possible de montrer qu'il existe α_1^* et α_2^*, $0 < \alpha_1^* < \alpha_2^*$, $\alpha_1^* + \alpha_2^* = \alpha$, tels que,

$$\mathbb{E}_{\mu,\sigma^2}[L_n(\widehat{s}_n^2, \alpha_1^*, \alpha_2^*)] \leq \mathbb{E}_{\mu,\sigma^2}[L_n(\widehat{s}_n^2, \alpha_1, \alpha_2)]$$

pour tout $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$ et tout (α_1, α_2) tels que $0 < \alpha_1 < \alpha_2$ et $\alpha_1 + \alpha_2 = \alpha$. On peut montrer que, lorsque n est grand, $\alpha_1^* \simeq \alpha_2^* \simeq \alpha/2$.

Exemple 6.4 (Région de confiance pour la moyenne et la variance)**:**
Supposons comme dans les exemples précédents que $X = (X_1, \ldots, X_n)$ est un n-échantillon i.i.d. d'une loi $\mathcal{N}(\mu, \sigma^2)$ mais cette fois nous cherchons à construire une *région de confiance* pour (μ, σ^2) de niveau de confiance $(1-\alpha)$. Notons les intervalles de confiance précédemment utilisé par :

$$I_1(X) = \left[\bar{X}_n - \widehat{s}_n t_{n-1}(1-\alpha/4)/\sqrt{n}, \bar{X}_n + \widehat{s}_n t_{n-1}(1-\alpha/4)/\sqrt{n}\right],$$

pour l'intervalle de confiance pour la moyenne μ de niveau de confiance $1-\alpha/2$ et :

$$I_2(X) = \left[\frac{(n-1)\widehat{s}_n^2}{x_{n-1}(1-\alpha/4)}, \frac{(n-1)\widehat{s}_n^2}{x_{n-1}(\alpha/4)}\right],$$

pour l'intervalle de confiance pour la variance σ^2 de niveau de confiance $1-\alpha/2$. Nous avons :

$$\mathbb{P}_{\mu,\sigma^2}\Big((\mu,\sigma^2) \in I_1(X) \times I_2(X)\Big) \geq 1 - \mathrm{P}_{\mu,\sigma^2}(\mu \notin I_1(X)) - \mathbb{P}_{\mu,\sigma^2}(\sigma^2 \notin I_2(X)) = 1-\alpha,$$

et donc $I(X) = I_1(X) \times I_2(X)$ est un intervalle de confiance de niveau de confiance supérieur à $(1-\alpha)$. Il est possible de montrer qu'en fait, le niveau de confiance de cet intervalle est exactement $(1-\alpha/2)^2$.

Dans certains cas, on est amené à déterminer des intervalles de confiance pour des fonctions d'un paramètre, soit $q(g(\theta))$, où q est une fonction monotone. Une façon simple, mais généralement sous-optimale, pour déterminer un tel intervalle est de remarquer que si $[m(X), M(X)]$ est un intervalle de confiance pour $g(\theta)$ de niveau de confiance $(1-\alpha)$, alors $q\left([m(X), M(X)]\right)$ est un intervalle de confiance de niveau $1-\alpha$ pour $q(g(\theta))$. Nous allons maintenant illustrer ce principe de calcul.

Exemple 6.5:
Soit $X_1, X_2, \ldots, X_n$ le nombre de minutes qu'un groupe d'utilisateurs tests d'Internet passent connectés par semaine. Nous modélisons ces v.a. par des v.a. i.i.d. de loi exponentielle de moyenne θ,

$$p_\theta(x) = \theta^{-1}\exp(x/\theta)\mathbb{1}(x \geq 0).$$

On cherche, pour un x donné, à construire un intervalle de niveau de confiance $1-\alpha$ pour la fonction

$$q(\theta) = \mathrm{P}_\theta([x,\infty)) = \mathbb{P}_\theta[X \geq x] = \exp(-x/\theta),$$

la probabilité que les utilisateurs-tests passent plus de x heures connectés dans la semaine. La variable $Z = \varphi(X,\theta) = 2n\bar{X}_n/\theta$ est distribuée suivant une loi χ^2_{2n} pour tout $\theta > 0$ et donc la variable Z peut être utilisée comme une variable pivotale. En utilisant Z comme pivot en résolvant en θ, nous obtenons l'intervalle de confiance de niveau $1-\alpha$:

$$m(X) = 2n\bar{X}_n/x_{2n}(1-\alpha/2) \leq \theta \leq 2n\bar{X}_n/x_{2n}(\alpha/2) = M(X),$$

où $x_{2n}(\beta)$ est le quantile d'ordre β de la loi χ^2_{2n}. $[q(m(X)), q(M(X))]$ est un intervalle de confiance de niveau $1-\alpha$ pour $q(\theta)$.

6.4 Dualité entre régions de confiance et tests d'hypothèse de base simple

Il existe des liens étroits entre tests statistiques et région de confiance. Ces liens peuvent être exploités pour construire des tests à partir de fonctions pivotales, ou construire des régions de confiance de niveaux donnés à partir d'une famille de tests.

Voici un exemple :

Exemple 6.6 (Test bilatéral pour la moyenne d'une gaussienne)**:**
Soit $(X_1, \dots, X_n)$ un n-échantillon i.i.d. d'une loi $\mathcal{N}(\mu, \sigma^2)$ où σ^2 est inconnue. Étant donné un $\mu_0 \in \mathbb{R}$, nous cherchons à tester $H_0 = \{\mu = \mu_0\}$ contre l'alternative $H_1 = \{\mu \neq \mu_0\}$.

Nous avons construit dans la partie précédente l'intervalle de confiance pour μ de niveau de confiance $(1-\alpha)$:

$$I_1(X) = \{x \ : \ |x - \bar{X}_n| \leq \widehat{s}_n t_{n-1}(1-\alpha/2)/\sqrt{n}\}.$$

Considérons la procédure de test suivante. Nous acceptons H_0 si

$$\mu_0 \in I_1(X) \ ,$$

et nous rejetons l'hypothèse dans le cas contraire. En notant $T = \sqrt{n}(\bar{X}_n - \mu_0)/\widehat{s}_n$, notre test accepte H_0 si $-t_{n-1}(1-\alpha/2) \leq T \leq t_{n-1}(1-\alpha/2)$. Ce test est bilatéral, car il rejette aussi bien les petites valeurs de T que les grandes valeurs de T. Contrairement aux tests unilatéraux considérés dans le paragraphe 5.5, ce test a de la puissance contre les alternatives où $\mu < \mu_0$ et $\mu > \mu_0$. L'erreur de première espèce est fixée égale à α. Nous avons ainsi, à partir d'un intervalle de confiance, construit une procédure de test.

Cet exemple est un cas particulier du principe de dualité entre intervalle de confiance et test. La région d'acceptation d'un test est un ensemble inclus dans $\mathcal{X}$, fixé, pour une hypothèse donnée sur le paramètre ; au contraire, une région de confiance est un ensemble de paramètres pour une observation donnée de loi inconnue.

Pour expliciter cette dualité, on va considérer une famille de tests dont les hypothèses de base dépendent d'un paramètre. On note $(H_0(t))_{t\in\Theta}$ la famille d'hypothèses de base simples :

$$H_0(t) : \theta = t \ . \tag{6.4}$$

Soit $\alpha \in (0,1)$. Pour tout $t \in \Theta$, on se donne une procédure de test $\delta(\cdot; t) : \mathcal{X} \to \{0,1\}$ de niveau α pour l'hypothèse $H_0(t)$. Notons $A(t) \subseteq \mathcal{X}$ la région d'acceptation de H_0 associée :

$$A(t) = \{x \in \mathcal{X} \ : \delta(x;t) = 0\} \ .$$

On définit maintenant la région de confiance *duale* $S(x) \subseteq \Theta$ associée à un valeur $x \in \mathcal{X}$ par

$$S(x) = \{t \in \Theta \ : x \in A(t)\} = \{t \in \Theta \ : \delta(x;t) = 0\} \ . \tag{6.5}$$

En d'autres termes, $S(x)$ est l'ensemble des θ que l'on aurait acceptés avec la procédure de test $\delta(\,\cdot\,, \theta)$, en ayant observé $X = x$. Formellement, S est l'« image réciproque » de x par A, au sens où l'on a la relation de dualité

$$\forall (x,t) \in \mathcal{X} \times \Theta, \quad x \in A(t) \iff t \in S(x) \ . \tag{6.6}$$

Alors, par définition, la probabilité de couverture de la région de confiance S est, pour tout $\theta \in \Theta$,

$$\begin{aligned}\mathbb{P}_\theta(\theta \in S(X)) &= \mathbb{P}_\theta(X \in A(\theta)) \\ &= \mathbb{P}_\theta\left(\delta(X;\theta) = 0\right) \\ &= 1 - \mathbb{P}_\theta\left[\delta(X;\theta) = 1\right] \\ &= 1 - \alpha,\end{aligned}$$

puisque l'on a choisi la procédure $\delta(\,\cdot\,,\theta)$ de telle manière que son risque de première espèce $\mathbb{P}_\theta(\delta(X,\theta) = 1) = \alpha$. Ceci étant vrai pour tout $\theta \in \Theta$, la région S est de niveau de confiance $1 - \alpha$.

Réciproquement, supposons maintenant que l'on dispose d'une région de confiance S' de niveau de confiance $1 - \alpha$ pour le paramètre θ. Pour tout $t \in \Theta$, soit $\delta'(\cdot;t)$ la procédure de test définie par

$$\delta'(x;t) = \begin{cases} 1 & \text{si } t \in S'(x) \\ 0 & \text{sinon} \end{cases} \tag{6.7}$$

En d'autres termes, après avoir observé x, la procédure de test $\delta(\,\cdot\,,\theta)$ accepte $H_0(\theta)$ si θ appartient à la région de confiance $S'(x)$. Alors, pour tout $\theta \in \Theta$, $\delta'(\cdot;\theta)$ est une procédure de test $\mathcal{X} \to \{0,1\}$ pour l'hypothèse $H_0(\theta)$ de niveau donné par

$$\mathbb{P}_\theta\left(\delta'(X;\theta) = 1\right) = 1 - \mathbb{P}_\theta\left(\theta \in S'(X)\right) = \alpha. \tag{6.8}$$

Ces relations de dualité entraînent le résultat suivant.

Théorème 6.4.1 (Dualité tests/régions de confiance)
Si pour tout $t \in \Theta$, $\delta(\cdot;t)$ est une procédure de test $\mathcal{X} \to \{0,1\}$ de niveau (risque de rejeter à tort) α pour l'hypothèse $H_0(t)$ définie par (6.4), alors la région de confiance S définie par (6.5) est de niveau de confiance $1 - \alpha$.

Si, de plus, pour tout $t \in \Theta$, $\delta(\cdot;t)$ est une procédure de test U.P.P. de niveau α pour l'hypothèse $H_0(t)$ contre l'hypothèse $H_1(t)$, alors, pour tout $t \in \Theta$, la région de confiance S minimise la probabilité $\mathbb{P}_\theta(t \in S(X))$ uniformément sur l'ensemble des $\theta \neq t$; autrement dit, pour tout $t \in \Theta$, toute région de confiance S' de niveau de confiance au moins égale à $1 - \alpha$ et tout $\theta \neq t$, on a

$$\mathbb{P}_\theta(t \in S(X)) \leq \mathbb{P}_\theta(t \in S'(X))\ . \tag{6.9}$$

Démonstration. La première partie du théorème a déjà été prouvée en (6.4). Il ne reste plus qu'à montrer (6.9). La probabilité de gauche dans cette équation s'écrit $\mathbb{P}_\theta(\delta(\cdot;t) = 0)$; c'est donc le risque de deuxième espèce de $\delta(\cdot;t)$ pris en θ (qui vérifie l'hypothèse $H_1(t)$). Puisque le test $\delta(\,\cdot\,,t)$ est supposé U.P.P. dans la classe des tests de niveau au plus α, il suffit donc de montrer que la probabilité de droite dans l'équation (6.9) est le risque de deuxième espèce d'une procédure de test de niveau au plus α. Pour cela il suffit de considérer le test $\delta'(\,\cdot\,;t)$ défini par (6.7) et d'utiliser la relation de dualité (6.6) entre δ' et S' qui implique que

$$\mathbb{P}_\theta(t \in S'(X)) = \mathbb{P}_\theta(X \in A(t)) = \mathbb{P}_\theta(\delta'(X;t) = 0) = R(\theta, \delta'(\,\cdot\,;t)).$$

■

6.5 Le cas du rapport de vraisemblance monotone

Sous l'hypothèse de rapport de vraisemblance monotone vue au chapitre 5, il est possible d'exploiter la dualité entre tests et région de confiance introduite au paragraphe 6.4 pour construire des intervalles de confiance et des bornes de confiance de niveaux de confiance donné.

Théorème 6.5.1
Soit $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$ un modèle paramétrique de paramètre scalaire, $\Theta \subseteq \mathbb{R}$, vérifiant l'hypothèse **(MON)** *pour la statistique $\mathbf{T} = T(X)$. Notons $F_\theta(z) = \mathbb{P}_\theta(\mathbf{T} \leq z)$. Supposons de plus que :*

(i) $z \to F_\theta(z)$ *est continue pour tout* $\theta \in \Theta$,

(ii) $\theta \to F_\theta(z)$ *est continue pour tout* $z \in T(\mathcal{X})$,

(iii) Pour tout $\alpha \in (0,1)$ *et* $z \in T(\mathcal{X})$, *l'équation en* θ *:* $F_\theta(z) = 1-\alpha$ *admet une solution unique.*

Notons $m_\alpha(z)$ la solution de l'équation $F_\theta(z) = \alpha$ et $M_\alpha(z)$ la solution de $F_\theta(z) = 1-\alpha$. Alors :

(1) $m_\alpha(\mathbf{T})$ *est une borne inférieure de confiance pour* θ *de niveau de confiance* $1-\alpha$ *: pour tout* $\theta \in \Theta$,
$$\mathbb{P}_\theta(m_\alpha(\mathbf{T}) \leq \theta) \geq 1-\alpha.$$

(2) $M_\alpha(\mathbf{T})$ *est une borne supérieure de confiance pour* θ *au niveau de confiance* $1-\alpha$ *: pour tout* $\theta \in \Theta$,
$$\mathbb{P}_\theta(M_\alpha(\mathbf{T}) \geq \theta) \geq 1-\alpha.$$

(3) *Pour tout* $\alpha_1, \alpha_2 \geq 0$ *tels que* $\alpha_1 + \alpha_2 < 1$, $[m_{\alpha_1}(\mathbf{T}), M_{\alpha_2}(\mathbf{T})]$ *est un intervalle de confiance pour* θ *de niveau* $1-(\alpha_1+\alpha_2)$ *: pour tout* $\theta \in \Theta$,
$$\mathbb{P}_\theta(m_{\alpha_1}(\mathbf{T}) \leq \theta \leq M_{\alpha_2}(\mathbf{T})) \geq 1-(\alpha_1+\alpha_2).$$

Démonstration. Considérons pour tout $\theta \in \Theta$ le test $\delta(\,\cdot\,;\theta)$ défini par :

$$\forall x \in \mathcal{X}, \quad \delta(x;\theta) = 0 \quad \Longleftrightarrow \quad T(x) > Q_\theta(\alpha),$$

où $Q_\theta(\alpha)$ est le quantile d'ordre α de la distribution F_θ. Nous avons sous les hypothèses énoncées dans le théorème, pour tout $\theta \in \Theta$:

$$\mathbb{P}_\theta\left(\delta(\,\cdot\,;\theta) = 0\right) = \mathbb{P}_\theta(\mathbf{T} > Q_\theta(\alpha)) = 1 - F_\theta\big[Q_\theta(\alpha)\big] = 1-\alpha.$$

Pour tout $t \in \Theta$, $\delta(\cdot, t)$ est donc un test de niveau α de l'hypothèse de base $H_0(t) = \{\theta = t\}$ (rappelons que, d'après le théorème 5.5.2, ce test est U.P.P. contre l'alternative composite $H_1(t) : \{\theta > t\}$).

Considérons maintenant la région duale $S(x)$, définie par

$$S(x) = \{\theta \in \Theta, \delta(x;\theta) = 0\} = \{\theta \in \Theta, T(x) \leq Q_\theta(1-\alpha)\}\,.$$

Le principe de dualité (Théorème 6.4.1) implique que l'ensemble $S(x)$ est une région de confiance de niveau $1-\alpha$ pour θ, puisque, pour tout $\theta \in \Theta$,

$$\mathbb{P}_t\left(t \in S(T)\right) = \mathbb{P}_t(T \leq Q_t(1-\alpha)) = 1-\alpha.$$

Il reste à prouver que la région de confiance $S(T)$ est ici un intervalle de la forme $[m_\alpha(T), \infty)$. En appliquant F_t aux deux membres de l'inégalité $y \leq Q_t(1-\alpha)$, nous avons :

$$S(y) = \{t \in \Theta, F_t(y) \leq 1-\alpha\}.$$

Soit $y \in T(\mathcal{X})$ et $\alpha \in (0,1)$ tels que $y = Q_t(1-\alpha)$, c'est-à-dire $F_t(y) = 1-\alpha$ ou encore $m_\alpha(y) = t$. Pour tout $t \in \Theta$, le test $\delta(\cdot; t)$ est U.P.P. contre les alternatives de la forme $H_1(t) : \{\theta < t\}$ dans la classe des tests de niveau au plus α. Ce test est en particulier plus puissant que le test de fonction critique constante α, ce qui implique que, pour $t' > t$,

$$\mathbb{P}_{t'}(T > y) = 1 - F_{t'}(y) \geq \alpha = 1 - F_t(y),$$

ce qui implique que $t \mapsto F_t(y)$ est une fonction *décroissante* de t. Par conséquent, la condition $F_t(y) \leq 1-\alpha$ équivaut à $t \geq m_\alpha(y)$, ce qui conclut la preuve de la première assertion. Les autres assertions se déduisent de la même façon. ■

Exemple 6.7 (Intervalle de confiance pour une loi de Poisson)**:**
Pour déterminer une borne maximale du degré de radioactivité d'une source, on enregistre les temps d'arrivée successifs de m particules sur un compteur. En supposant que le radionucléide se décompose en émettant des particules suivant une loi de Poisson, les temps d'arrivée T_i des particules sur le compteur sont distribuées suivant une loi exponentielle de paramètre θ, où θ est l'intensité du processus :

$$p_\theta(t_1, \ldots, t_m) = \theta^m e^{-\theta \sum_{i=1}^m t_i}, \quad t_1, \ldots, t_m \geq 0.$$

Notons $T = \sum_{i=1}^m T_i$ la durée totale d'observation. La variable $2\theta T$ est distribuée suivant une loi de χ^2 à $2m$ degrés de liberté. La région d'acceptation du test $H_0(t) = \{\theta = t\}$ contre $H_1(t) = \{\theta < t\}$ est de la forme $2tT \leq x_{2m,\alpha}$, où $x_{2m,\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi χ^2_{2m} (loi du χ^2 à $2m$ degrés de liberté). L'ensemble $S(t_1, \ldots, t_m)$ est donc l'ensemble des t vérifiant, $t \leq x_{2m,\alpha}/2T$, et le théorème précédent montre que $M(T) = x_{2m,\alpha}/2T$ est une borne supérieure de confiance pour le paramètre θ de niveau de confiance $1-\alpha$.

Annexe A

Rappels de probabilité

A.1 Espace de probabilité

Soit un espace abstrait Ω, appelé *espace des épreuves*. Un élément ω de Ω est appelé une *épreuve* ou *réalisation* : ω correspond au résultat d'une expérience aléatoire. L'ensemble Ω est souvent appelé l'ensemble des *épreuves* ou des *réalisations*. L'espace Ω dépend bien entendu de l'expérience aléatoire que l'on cherche à modéliser. Nous verrons des exemples dans la suite. Nous construisons sur cet ensemble d'épreuves un ensemble de parties $\mathcal{F}$, muni d'une structure de *tribu*

Définition A.1.1 (Tribu)**.** *Une tribu $\mathcal{F}$ est un ensemble de parties de Ω vérifiant les propriétés suivantes :*

1. $\Omega \in \mathcal{F}$,
2. *si $A \in \mathcal{F}$, alors $A^c \in \mathcal{F}$, où A^c est le complémentaire de A, $A^c := \Omega \backslash A = \{x \in \Omega, x \notin A\}$ ("stabilité par passage au complémentaire").*
3. *si $(A_n, n \in \mathbb{N})$ est une suite de parties de Ω, alors, $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{F}$ ("stabilité par réunion dénombrable").*

Un élément d'une tribu s'appelle un *événement* (en théorie de la mesure, de tels éléments sont appelés *ensembles mesurables*). Deux événements A et B sont dits *incompatibles*, si $A \cap B = \emptyset$. L'ensemble vide est appelé l'événement *impossible*. A l'inverse, Ω est l'événement *certain*. Le couple $(\Omega, \mathcal{F})$ constitué d'un ensemble d'épreuves et d'une tribu d'événements est un *espace probabilisable*. L'ensemble des parties de Ω, $\mathcal{P}(\Omega)$ est une tribu. Toutes les tribus définies sur Ω sont des sous-ensembles de $\mathcal{P}(\Omega)$. L'ensemble $\{\emptyset, \Omega\}$ est aussi une tribu. Cette tribu est contenue dans toutes les tribus définies sur Ω. L'intersection d'une famille quelconque de tribus est encore une tribu.

Définition A.1.2 (Tribu engendrée, $\sigma(\mathcal{A})$)**.** *La tribu engendrée par une classe de parties $\mathcal{A}$ de Ω, notée $\sigma(\mathcal{A})$ est la plus petite tribu contenant $\mathcal{A}$.*

La tribu engendrée $\sigma(\mathcal{A})$ est l'intersection de toutes les tribus contenant $\mathcal{A}$. Notons que toute classe $\mathcal{A}$ est incluse dans $\mathcal{P}(\Omega)$, et donc qu'il existe toujours au moins une tribu contenant $\mathcal{A}$. La notion de *tribu borélienne* est liée à la structure "topologique" de l'ensemble de base : c'est la tribu engendrée par l'ensemble des ouverts de la topologie. Nous considérerons dans ce chapitre uniquement la tribu borélienne de $\mathbb{R}^d$, en commençant par le cas le plus simple de la droite réelle $\mathbb{R}$.

Définition A.1.3 (Tribu borélienne)**.** *La tribu borélienne ou tribu de Borel de $\mathbb{R}$ est la tribu engendrée par la classe des intervalles ouverts. On la note $\mathcal{B}(\mathbb{R})$. Un élément de cette tribu est appelé une partie borélienne ou un borélien.*

Tout intervalle ouvert, fermé, semi-ouvert, appartient à $\mathcal{B}(\mathbb{R})$. Il en est de même de toute réunion finie ou dénombrable d'intervalles (ouverts, fermés, ou semi-ouverts). La tribu $\mathcal{B}(\mathbb{R})$ est aussi la tribu engendrée par l'une quelconque des quatre classes suivantes d'ensembles :

$$\mathcal{I} = \{]-\infty, x], x \in \mathbb{R}\}\ \mathcal{I}' = \{]-\infty, x]; x \in \mathbb{Q}\},$$
$$\mathcal{J} = \{]-\infty, x[, x \in \mathbb{R}\}\ \mathcal{J}' = \{]-\infty, x[; x \in \mathbb{Q}\}.$$

De façon similaire, la tribu borélienne $\mathcal{B}(\mathbb{R}^d)$ de $\mathbb{R}^d$ est la tribu engendrée par les rectangles ouverts $\prod_{i=1}^{d}]a_i, b_i[$. Le théorème suivant sera d'un usage constant dans la suite.

Théorème A.1.4 (Classe monotone)
Soient $\mathcal{C} \subset \mathcal{M} \subset \mathcal{P}(\Omega)$. On suppose que :
- *$\Omega \in \mathcal{M}$,*
- *pour tout $A, B \in \mathcal{M}$, $A \subset B$ implique que $B \setminus A \in \mathcal{M}$,*
- *$\mathcal{M}$ est stable par limite croissante.*

Alors, $\sigma(\mathcal{C}) \subset \mathcal{M}$.

A.2 Probabilité

Définition A.2.1 (Probabilité)**.** *On appelle probabilité sur $(\Omega, \mathcal{F})$, une application $\mathbb{P} : \mathcal{F} \mapsto [0, 1]$, qui vérifie les propriétés suivantes :*

1. *$\mathbb{P}(\Omega) = 1$,*
2. *("σ-additivité") si $(A_n, n \in \mathbb{N})$ est une suite d'éléments de $\mathcal{F}$ deux à deux disjoints, (*i.e.*$A_i \cap A_j = \emptyset$ pour $i \neq j$), alors :*

$$\mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_i\right) = \sum_{i=0}^{\infty} \mathbb{P}(A_i).$$

On vérifie aisément les propriétés suivantes : A_n,A et B étant des événements :

$$A \subset B,\ \mathbb{P}(A) \leq \mathbb{P}(B),\ \mathbb{P}(A^c) = 1 - \mathbb{P}(A),$$
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B),$$
$$A_n \nearrow B,\ \mathbb{P}(A_n) \nearrow \mathbb{P}(A),\ A_n \searrow A,\ \mathbb{P}(A_n) \searrow \mathbb{P}(A),\ \mathbb{P}(\bigcup_n A_n) \leq \sum_n \mathbb{P}(A_n).$$

Définition A.2.2 (Ensemble négligeable)**.** *On dit qu'un ensemble $A \subset \Omega$ est $\mathbb{P}$-négligeable (ou plus simplement négligeable, s'il n'y a pas d'ambiguïté sur la mesure de probabilité) s'il existe un ensemble $B \in \mathcal{F}$, tel que $A \subset B$ et $\mathbb{P}(B) = 0$.*

Remarquons que les ensembles négligeables ne sont pas nécessairement des éléments de la tribu $\mathcal{F}$. Une propriété est dite $\mathbb{P}$-presque sûre, si la propriété est vérifiée sur un ensemble dont le complémentaire est $\mathbb{P}$-négligeable.

Définition A.2.3 (Espace de probabilité)**.** *Le triplet* $(\Omega, \mathcal{F}, \mathbb{P})$ *définit un* espace de probabilité*.*

Définition A.2.4 (Tribu complète)**.** *On dira que la tribu* $\mathcal{F}$ *est* complète *si tous les ensembles négligeables de* Ω *sont éléments de* $\mathcal{F}$*.*

Il est facile de construire une tribu $\mathcal{F}'$ qui contient $\mathcal{F}$ et d'étendre $\mathbb{P}$ à $\mathcal{F}'$ de telle sorte que $\mathcal{F}'$ soit complète pour l'extension de $\mathbb{P}$. Pour éviter des complications techniques inutiles, nous supposerons désormais que toutes les tribus que nous manipulerons sont complètes. Rappelons pour conclure ce paragraphe deux résultats techniques d'usage constant.

Définition A.2.5 (π–système)**.** *On appelle un* π–système *une famille d'ensembles stable par intersection finie.*

Théorème A.2.6 (π–système)
Soient μ *et* ν *deux mesures sur* $(\Omega, \mathcal{F})$ *et soit* $\mathcal{C} \subset \mathcal{F}$ *un* π*-système. On suppose que pour tout* $C \in \mathcal{C}$*,* $\mu(C) = \nu(C) < \infty$ *et que* $\mu(\Omega) = \nu(\Omega) < \infty$*. Alors* $\mu(A) = \nu(A)$ *pour tout* $A \in \sigma(\mathcal{C})$*.*

Soit E un ensemble.

Définition A.2.7 (Algèbre)**.** *Une famille* $\mathcal{E}_0$ *de sous-ensembles de* E *est appelé une* algèbre *si* ***(i)*** $E \in \mathcal{E}_0$*,* ***(ii)*** $F \in \mathcal{E}_0 \Longrightarrow F^c \in \mathcal{E}_0$ *et* ***(iii)*** $F, G \in \mathcal{E}_0 \Longrightarrow F \cup G \in \mathcal{E}_0$*.*

A la différence des tribus, nous ne supposons pour les algèbres que la stabilité par union finie (et non infinie dénombrable). Une fonction d'ensembles μ définie sur $\mathcal{E}_0$ est dite σ-additive, si pour toute union dénombrable d'éléments $F_i \in \mathcal{E}_0$, $F_i \cap F_j = \emptyset$, telle que $\bigcup_i F_i \in \mathcal{E}_0$, $\mu\left(\bigcup_i F_i\right) = \sum_i \mu(F_i)$.

Théorème A.2.8 (Théorème d'extension de Carathéodory)
Soit E *un ensemble et* $\mathcal{E}_0$ *une algèbre sur* E*. Soit* μ_0 *une fonction d'ensembles* σ*-additive, telle que* $\mu_0(E) < \infty$*. Alors, il existe une unique mesure* μ *sur* $\mathcal{E} := \sigma(\mathcal{E}_0)$ *telle que* $\mu = \mu_0$ *sur* $\mathcal{E}_0$*.*

Exemple A.1:
Pour illustrer l'utilisation de ce théorème, rappelons la construction de la mesure de Lebesgue. Soit $\mathcal{C}$ l'ensemble des parties de $[0, 1]$ pouvant s'écrire sous la forme d'une union finie d'intervalles ouverts à gauche et fermés à droite, *i.e.*$F \in \mathcal{C}$ si :

$$F =]a_1, b_1] \cup \cdots \cup]a_r, b_r].$$

On vérifie facilement que $\mathcal{C}$ est une algèbre. La tribu engendrée par $\mathcal{C}$, $\sigma(\mathcal{C}) = \mathcal{B}([0, 1])$, est la tribu borélienne sur $[0, 1]$. Pour $F \in \mathcal{F}_0$ considérons :

$$\lambda_0(F) = \sum_i (b_i - a_i).$$

On vérifie que λ_0 est une fonction positive et additive. On peut démontrer que λ_0 est σ-additive, *i.e.*pour toute union dénombrable d'ensembles $F_i \in \mathcal{F}_0$ disjoints 2 à 2 tels que $\bigcup_i F_i \in \mathcal{F}_0$, $\lambda_0(F) = \sum_i \lambda_0(F_i)$ (cette partie de la preuve n'est pas immédiate). Le théorème de Carathéodory permet de montrer que λ_0 a une extension unique λ sur $\mathcal{B}([0, 1])$, appelée *mesure de Lebesgue* sur $[0, 1]$.

A.3 Variables aléatoires

Définition

Soient Ω et E deux ensembles munis respectivement des tribus $\mathcal{F}$ et $\mathcal{E}$. Soit f une application d'un espace Ω dans un espace E. L'*image réciproque* d'une partie A de E par f est la partie de Ω notée $f^{-1}(A)$ définie par :

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

Les propriétés suivantes, où A et les ensembles A_i sont des parties quelconques de F et I est un ensemble fini, dénombrable, ou infini non dénombrable, se vérifient immédiatement :

$$\begin{aligned} &f^{-1}(E) = \Omega, \quad f^{-1}(\emptyset) = \emptyset, \quad f^{-1}(A^c) = \left(f^{-1}(A)\right)^c, \\ &f^{-1}\left(\bigcup_{i\in I} A_i\right) = \bigcup_{i\in I} f^{-1}(A_i), \quad f^{-1}\left(\bigcap_{i\in I} A_i\right) = \bigcap_{i\in I} f^{-1}(A_i). \end{aligned} \tag{A.2}$$

Si $\mathcal{A}$ est une classe quelconque de parties de E, on note $f^{-1}(\mathcal{A})$ la classe de parties de Ω définie par : $f^{-1}(\mathcal{A}) = \{f^{-1}(A) : A \in \mathcal{A}\}$. Il découle immédiatement des propriétés précédentes que si $\mathcal{E}$ est une tribu de E, $f^{-1}(\mathcal{E})$ est une tribu de Ω.

Définition A.3.1 (Variable aléatoire, v.a.)**.** *Soient $(\Omega, \mathcal{F})$ et $(E, \mathcal{E})$ deux espaces probabilisables, et X une application de Ω dans E. On dit que X est une v.a. de $(\Omega, \mathcal{F})$ dans $(E, \mathcal{E})$ si la tribu $X^{-1}(\mathcal{E})$ est contenue dans $\mathcal{F}$, ce qui revient à dire que $X^{-1}(A) \in \mathcal{F}$ pour tout ensemble $A \in \mathcal{E}$.*

Lorsque le cardinal de l'ensemble E est fini ou dénombrable, la tribu $\mathcal{E}$ est le plus souvent choisie comme l'ensemble des parties de E, $\mathcal{E} = \mathcal{P}(E)$, et une v.a. X définie sur $(\Omega, \mathcal{F})$ à valeurs dans $(E, \mathcal{E})$ est dite *discrète*. Lorsque $E = \bar{\mathbb{R}}^+$ (où $\bar{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ et $\mathbb{R}^+$ l'ensemble des réels positifs) et $\mathcal{E} = \mathcal{B}(\bar{\mathbb{R}}^+)$ est la tribu borélienne de $\bar{\mathbb{R}}^+$, on dit que X est une v.a. positive. Si $E = \bar{\mathbb{R}}$ et $\mathcal{E} = \mathcal{B}(\bar{\mathbb{R}})$, on dit que X est une v.a. réelle. Si $E = \bar{\mathbb{R}}^d$ et $\mathcal{E} = \mathcal{B}(\bar{\mathbb{R}}^d)$, on dit que X est une variable vectorielle (ou vecteur aléatoire). Soit $(X_i, i \in I)$ une famille de v.a. à valeurs dans $(E, \mathcal{E})$ (I étant un ensemble quelconque, non nécessairement dénombrable).

Définition A.3.2 (Tribu engendrée par une famille de v.a.)**.** *On appelle **tribu engendrée** par $(X_i, i \in I)$ et on note $\sigma(X_i,\, i \in I)$ la plus petite tribu $\mathcal{G}$ de Ω qui soit telle que toutes les v.a. X_i soient $\mathcal{G}$-mesurable.*

A titre d'illustration, soit une v.a. à valeur dans $(E, \mathcal{E})$. La tribu $\sigma(X)$ est la tribu engendrée par la classe d'ensembles $X^{-1}(B) := \{\omega : X(\omega) \in B\}$, où B parcourt $\mathcal{E}$. Comme $X^{-1}(\mathcal{E}) := \{Y^{-1}(B) : B \in \mathcal{E}\}$ est une tribu, on a :

$$\sigma(X) := \sigma\left(X^{-1}(B)\,,\, B \in \mathcal{E}\right) = Y^{-1}(\mathcal{E}).$$

Le résultat suivant est important car il donne une description simple des v.a. $\sigma(X)$-mesurables.

Théorème A.3.3
Soit X une v.a. à valeur dans $(E, \mathcal{E})$. Toute v.a. réelle Y est $\sigma(X)$-mesurable si et seulement s'il existe une fonction mesurable $f : E \to \mathbb{R}$ telle que $Y = f(X)$.

DÉMONSTRATION. Pour tout f mesurable, il est clair que $f(X)$ est $\sigma(X)$-mesurable. La réciproque est laissée à titre d'exercice : elle nécessite un résultat d'approximation des v.a. positives par des variables étagées introduit ultérieurement (lemme A.3.10). ■

Définition A.3.4 (limite inférieure et limite supérieure)**.** *Soit $\{X_n\}$ une suite de v.a. de $(\Omega, \mathcal{F}) \mapsto (\bar{\mathbb{R}}, \mathcal{B}(\mathbb{R}))$. On appelle limite supérieure et limite inférieure de la suite de v.a. $\{X_n\}_{n\geq 1}$ les applications suivantes :*

$$\begin{aligned} \text{limsup}_n X_n(\omega) &= \lim_n \searrow \sup_{m\geq n} X_m(\omega) = \inf_n \sup_{m\geq n} X_m(\omega), \\ \text{liminf}_n X_n(\omega) &= \lim_n \searrow \inf_{m\geq n} X_m(\omega) = \sup_n \inf_{m\geq n} X_m(\omega). \end{aligned} \tag{A.3}$$

Notons que les applications $\text{limsup}_n X_n$ et $\liminf_n X_n$ définies ci-dessus sont a-priori à valeurs dans $\bar{\mathbb{R}}$ même si les v.a. X_n sont à valeurs dans $\mathbb{R}$.

Proposition A.3.5
Soit $\{X_n\}_{n\in\mathbb{N}}$ une suite de v.a. sur $(\Omega, \mathcal{F})$ à valeurs dans $(\bar{\mathbb{R}}, \mathcal{B}(\bar{\mathbb{R}}))$.
- $\sup_n X_n$ *et* $\inf_n X_n$ *sont des v.a.,*
- $\text{limsup}_n X_n$ *et* $\text{liminf}_n X_n$ *sont des v.a.,*
- *L'ensemble* $\{\omega \in \Omega : \limsup_n X_n(\omega) = \text{liminf}_n X_n(\omega)\}$ *est élément de la tribu* $\mathcal{F}$.

DÉMONSTRATION. Pour (a), on utilise le fait que $\{\sup_n X_n \leq x\} = \bigcap_n \{X_n \leq x\}$ et $\{\inf_n X_n < x\} = \bigcup_n \{X_n < x\}$. (b) s'obtient par application répétée de (a). Notons $Y = \text{limsup}_n X_n$ et $Z = \text{liminf}_n X_n$. L'ensemble des épreuves ω pour lesquels la suite $\{X_n(\omega)\}_{n\in\mathbb{N}}$ admet une limite est par définition égal à $\{Y - Z = 0\}$. Comme Y et Z sont des v.a., $Y - Z$ est une v.a., ce qui conclut la preuve. ■

Espérance d'une variable aléatoire

Nous rappelons succinctement dans le paragraphe suivant quelques éléments de théorie d'intégration.

Définition A.3.6 (v.a. étagée)**.** *On dit qu'une v.a. X définie sur $(\Omega, \mathcal{F})$ et à valeurs dans $(\bar{\mathbb{R}}, \mathcal{B}(\bar{\mathbb{R}}))$ est étagée si elle ne prend qu'un nombre **fini** de valeurs dans $\bar{\mathbb{R}}$.*

On note dans la suite $e\mathcal{F}^+$ l'ensemble des v.a. étagées positives. Cet ensemble n'est pas un espace vectoriel, mais il est stable par addition et par multiplication par les réels positifs ($e\mathcal{F}^+$ est un cône). Etant données des nombres $a_1, \ldots, a_n$ de $\mathbb{R}^+$ et des ensembles $A_1, \ldots, A_n \in \mathcal{F}$, on obtient une v.a. positive $X \in e\mathcal{F}^+$ en posant :

$$X = \sum_{k=1}^{n} a_k \mathbb{1}(A_k), \quad A_k \in \mathcal{F}, \tag{A.4}$$

où $\mathbb{1}(A)$ est la fonction indicatrice de A, c'est-à-dire la fonction $\Omega \to \{0, 1\}$ définie en tout $\omega \in \Omega$ par

$$\mathbb{1}_A(\omega) = \begin{cases} 1 & \omega \in A, \\ 0 & \omega \notin A \end{cases} \tag{A.5}$$

Il est clair que cette fonction ne peut prendre qu'un nombre fini de valeurs, qui sont les sommes d'un nombre quelconque de a_i. Il y a évidemment de multiples façons d'écrire (A.4).

Inversement, toute v.a. $X \in e\mathcal{F}^+$ s'écrit sous la forme (A.4) et même admet une écriture (A.4) *canonique* qui est unique. Soit $\mathcal{X}$ l'ensemble des valeurs prises par X, et soit pour $a \in \mathcal{X}$, $A_a = X^{-1}(\{a\})$. Les ensembles $A_a \in \mathcal{F}$ constitue une partition finie de Ω et on a :

$$X = \sum_{a \in \mathcal{X}} a\mathbb{1}(A_a). \tag{A.6}$$

Définition A.3.7 (Espérance d'une v.a. étagée positive)**.** *Soit $\mathbb{P}$ une probabilité sur $(\Omega, \mathcal{F})$. On appelle* ***espérance*** *par rapport à la probabilité $\mathbb{P}$ de la v.a. étagée X admettant la décomposition canonique (A.6) et on note $\mathbb{E}[X]$ le nombre de $\bar{\mathbb{R}}^+$ suivant*

$$\mathbb{E}[X] = \sum_{a \in \mathcal{X}} a\mathbb{P}[A_a].$$

L'intégrale de la v.a. constante $X = a \geq 0$ vaut a. Si $A \in \mathcal{F}$, l'espérance de la v.a. $X = \mathbb{1}(A)$ vaut $\mathbb{P}(A)$. La proposition suivante découle de façon immédiate de la construction précédente.

Proposition A.3.8
Soient X, Y deux éléments de $e\mathcal{F}^+$. Alors pour $a, b \geq 0$, $aX + bY \in e\mathcal{F}+$ et

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

Si $X \leq Y$, alors $\mathbb{E}[X] \leq \mathbb{E}[Y]$.

Le résultat technique suivant est la clef de voûte de la construction

Lemme A.3.9
Soient $X_n, Y_n \in e\mathcal{F}$ deux suites croissantes telles que $\lim \nearrow X_n = \lim \nearrow Y_n$. *Alors,* $\lim \nearrow \mathbb{E}[X_n] = \lim \nearrow \mathbb{E}[Y_n]$.

Notons $\mathcal{F}^+$ l'ensemble des v.a. positives. Soit $X \in \mathcal{F}^+$. Le résultat suivant est à la base de la construction de l'intégrale

Lemme A.3.10
Toute v.a. X positive est limite d'une suite croissante de fonctions étagées.

Il suffit de considérer la suite :

$$X_n(\omega) = \sum_{k=0}^{n2^n-1} \frac{k}{2^n} \mathbb{1}(k/2^n \leq X(\omega) \leq (k+1)/2^n) + n\mathbb{1}(X(\omega) \geq n$$

Le lemme A.3.10 montre qu'il existe une suite $X_n \in e\mathcal{F}$ telle que $X_n \nearrow X$; la monotonicité de l'espérance assure que $\mathbb{E}[X_n]$ est une suite croissante, et donc que cette suite a une limite α. Le lemme A.3.9 montre que cette limite ne dépend pas du choix de la suite $\{X_n\}$. On a en particulier :

$$\alpha = \lim \nearrow \sum_{k=0}^{n2^n} \frac{k}{2^n} \mathbb{P}(\{\omega : k/2^n \leq X(\omega) < (k+1)/2^n\}) + n\mathbb{P}(\{\omega : X(\omega) \geq n\}).$$

Définition A.3.11 (Espérance d'une v.a. positive)**.** *Soit X une v.a. positive. On appelle* ***espérance*** *de X par rapport à la probabilité $\mathbb{P}$ le nombre suivant de $[0,\infty]$:*

$$\mathbb{E}[X] = \int X \, d\mathbb{P} = \lim_n \nearrow \mathbb{E}[X_n],$$

où $\{X_n\}$ est une suite croissante de v.a. étagées telle que $\lim_n \nearrow X_n = X$.

On montre aisément que :

$$\mathbb{E}[X] = \sup_{Y \in e\mathcal{F}^+ Y \leq X} \mathbb{E}[Y],$$

cette dernière relation étant souvent utilisée comme définition de l'espérance. Nous pouvons maintenant énoncé l'un des résultats essentiel de la théorie :

Théorème A.3.12(i) *Si $(a,b) \in \mathbb{R}^+$, et $X, Y \in \mathcal{F}^+$, on a :*

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

(ii) *Si $X, Y \in \mathcal{F}^+$ et si $X \leq Y$, on a $\mathbb{E}[X] \leq \mathbb{E}[Y]$.*

Théorème de convergence monotone *Soit $\{X_n\}_{n\in\mathbb{N}}$ une suite croissante de v.a. de $\mathcal{F}^+$ et soit $X = \lim_n X_n$. Alors $\lim_n \mathbb{E}[X_n] = \mathbb{E}[X]$.*

Lemme de Fatou *Soit $\{X_n\}$ est une suite de v.a. de $\mathcal{F}^+$. Alors, :*

$$\mathbb{E}\left[\text{liminf}_n X_n\right] \leq \text{liminf}_n \mathbb{E}[X_n].$$

Il nous reste à définir l'espérance des v.a. réelles de signe quelconque. Pour cela, on utilise le fait qu'une v.a. réelle est toujours la différence de deux v.a. positives, cette décomposition n'étant bien sûr pas unique. Nous utilisons dans la suite la décomposition canonique en partie positive et partie négative, qui sont les v.a. définies par :

$$X^+ := X \wedge 0 \quad \text{et} \quad X^- := (-X) \wedge 0,$$

où $a \wedge b = \max(a,b)$. On vérifie aisément que $X = X^+ - X^-$ et $|X| = X^+ + X^-$. Cette décomposition est minimale dans le sens où, pour toute autre décomposition de X de la forme $X = Y - Z$ avec $Y \in \mathcal{F}^+$ et $Z \in \mathcal{F}^+$, nous avons $Y \geq X^+$ et $Z \geq X^-$.

Définition A.3.13 (Espérance, v.a. intégrable)**.** *On dit que la v.a. X est* ***intégrable*** *si $\mathbb{E}\left[|X|\right] < \infty$, ce qui équivaut à $\mathbb{E}[X^+] < \infty$ et $\mathbb{E}[X^-] < \infty$. Dans un tel cas, on appelle* ***espérance*** *de X par rapport à la probabilité $\mathbb{P}$ le nombre de $[0,\infty[$:*

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-].$$

On pose $\mathcal{F}$ l'espace des v.a. intégrables :

$$\mathcal{L}^1 = \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P}) = \{X \in f\mathcal{F}, \mathbb{E}\left[|X|\right] < \infty\}$$

Il est facile de voir que $\mathcal{L}^1$ est un espace vectoriel (car $|X+Y| \leq |X| + |Y|$, et par monotonicité de l'espérance) et que $X \mapsto \mathbb{E}[X]$ est une forme linéaire positive. De plus, pour $X \in \mathcal{L}^1$, $|\mathbb{E}[X]| \leq \mathbb{E}|X|$. Les propriétés suivantes découlent directement des théorèmes classiques de la théorie de la mesure (à savoir, le lemme de Fatou, et le théorème de convergence dominée)

Proposition A.3.14 — *("Lemme de Fatou") Si $X_n \geq 0$, alors $\mathbb{E}[\liminf X_n] \leq \liminf \mathbb{E}[X_n]$,*
— *("Convergence dominée") Si, pour tout $n \geq 1$, $|X_n(\omega)| \leq Y(\omega)$, $\mathbb{P}$-ps, et $Y \in \mathcal{L}^1$, alors $\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X]$*

Nous utiliserons de façon très fréquente dans la suite les résultats ci-dessus ; nous donnons toutefois sans attendre quelques exemples d'applications utiles :

Exemple A.2: — Soit (Z_k) une suite de v.a. positives. Alors $\mathbb{E}\left[\sum_k Z_k\right] = \sum \mathbb{E}[Z_k] \leq \infty$ (application de la convergence monotone et de la linéarité de l'espérance).
— Soit (Z_k) une suite de v.a. positives, telle que $\sum \mathbb{E}[Z_k] < \infty$. Alors $\sum Z_k$ est fini p.s. et donc $Z_k \to 0$ p.s.

Nous admettrons le résultat suivant (cf. le cours d'intégration)

Théorème A.3.15
*Soit X une v.a. de $(\Omega, \mathcal{F})$ dans $(E, \mathcal{E})$ et $\mathbb{P}$ une probabilité sur $(\Omega, \mathcal{F})$. La formule $\mathbb{P}_X(A) := \mathbb{P}(X^{-1}(A))$ définit une probabilité sur $(E, \mathcal{E})$, appelée **probabilité image** de $\mathbb{P}$ par X. Cette probabilité vérifie, pour toute fonction f positive mesurable :*

$$\int f \circ X(\omega)\, d\mathbb{P}(\omega) = \int f(x)\, d\mathbb{P}_X(x)$$

Définition A.3.16 (Loi d'une variable aléatoire)**.** *On appelle* loi de X *la probabilité image de $\mathbb{P}$ par X.*

La loi d'une variable aléatoire réelle est donc une probabilité sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Il est souvent pratique de spécifier la loi de probabilité d'une variable aléatoire réelle par la donnée de sa fonction de répartition,

Définition A.3.17 (Fonction de répartition)**.** *La fonction de répartition de la v.a. réelle X est la fonction $F_X : \mathbb{R} \mapsto [0, 1]$, définie par :*

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

Si X est une v.a. à valeurs dans $\mathbb{R}^d$, sa fonction de répartition est définie sur $\mathbb{R}^d$ par

$$F_X(\mathbf{x}) = \mathbb{P}_X\left(\prod_{k=1}^{d}]-\infty, x_k]\right) = \mathbb{P}(X \leq \mathbf{x}), \quad \mathbf{x} = (x_1, \ldots, x_d).$$

La fonction de répartition est une fonction croissante, continue à droite : on remarque en effet que $]-\infty, x] = \bigcap]-\infty, x_n]$, pour toute suite décroissante x_n, telle que $\lim_{n\to\infty} x_n = x$. La σ-additivité impose donc que $F_X(x) = \lim_{n\to\infty} F(x_n)$, et donc plus généralement que $\lim_{h\to 0+} F_X(x+h) = F_X(x)$. Un raisonnement similaire montre que la fonction de répartition admet en chaque point une limite à gauche : $\lim_{h\to 0-} F_X(x+h) = \mathbb{P}_X(]-\infty, x[) = F_X(x-)$. Remarquons aussi que :

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

La fonction de répartition F_X caractérise la loi $\mathbb{P}_X$, puisque pour tout intervalle $]a, b]$ $(b > a)$, on a $\mathbb{P}_X(]a, b]) = F_X(b) - F_X(a)$ et qu'une mesure borélienne sur $\mathbb{R}$ est déterminée de façon unique par la donnée des mesures de tels intervalles (cf. Livre A, chapitre 2).

A.4 Quelques inégalités utiles

L'inégalité élémentaire suivante joue un rôle fondamental.

Proposition A.4.1 (Inégalité de Markov)
Soit Z une v.a et $g : \mathbb{R} \mapsto [0, \infty]$ une fonction borélienne croissante. Alors :

$$\mathbb{E}[g(Z)] \geq \mathbb{E}[g(Z)\mathbb{1}(Z \geq c)] \geq g(c)\mathbb{P}[Z \geq c].$$

En particulier, on a :

Corollaire A.4.2 (Inégalité de Bienaymé–Tchebychev)
Soit Z une v.a. à valeurs dans $\mathbb{R}^d$ vérifiant $\mathbb{E}[\|Z\|^p] < \infty$ pour un $p > 0$. Alors, pour tout $\delta > 0$,

$$\mathbb{P}(\|Z\| > \delta) \leq \mathbb{E}[\|Z\|^p]\delta^{-p} .$$

Une fonction $c : G \mapsto \mathbb{R}$ où G est un intervalle ouvert de $\mathbb{R}$ est dite *convexe* si, pour tout $x, y \in G$ et tout p, q, $p + q = 1$:

$$c(px + qy) \leq pc(x) + qc(y).$$

A titre d'exemples, les fonctions $|x|$, x^2, $e^{\theta x}$ sont des fonctions convexes. La proposition suivante est souvent utile.

Proposition A.4.3 (Inégalité de Jensen)
Soit $g : I \mapsto \mathbb{R}$ une fonction convexe sur un intervalle ouvert I de $\mathbb{R}$ et soit X une v.a. réelle vérifiant les propriétés suivantes :

$$\mathbb{P}[X \in G] = 1, \ \mathbb{E}\left[|g(X)|\right] < \infty$$

Alors, $g(\mathbb{E}[X]) \leq \mathbb{E}\left[g(X)\right]$.

Proposition A.4.4 (Inégalité de Cauchy–Schwarz)
Soient Y et Z deux v.a. à valeurs dans $\mathbb{R}$. On a

$$(\mathbb{E}[YZ])^2 \leq \mathbb{E}[Y^2]\,\mathbb{E}[Z^2] \ ;,$$

avec égalité si et seulement si Y et Z sont co–linéaires : il existe $\lambda \in \mathbb{R}$ tel que $Y + \lambda Z = 0$ p.s.

Corollaire A.4.5 (Inégalité de Bienaymé–Cantelli)
Soit Z une v.a. à valeurs dans $\mathbb{R}$ vérifiant $\mathbb{E}[Z^2] < \infty$. Alors, pour tout $\delta > 0$,

$$\mathbb{P}(Z > \delta) \leq \frac{\mathbb{V}\mathrm{ar}(Z)}{\mathbb{V}\mathrm{ar}(Z) + \delta^2} .$$

A.5 Mesures σ-finies

Soit $(\Omega, \mathcal{F})$ un espace mesurable. On rappelle qu'une mesure positive μ est une application $\mathcal{F} \to [0, \infty]$ σ-additive telle que $\mu(\emptyset) = 0$. Soient μ et ν deux mesures positives sur $(\Omega, \mathcal{F})$ et

f une fonction mesurable de Ω dans $\mathbb{R}_+$. On dit que μ a une **densité** f par rapport à ν, si, pour tout $A \in \mathcal{F}$,

$$\mu(A) = \int_A f \, d\nu.$$

Cette fonction f est unique dans le sens où, s'il existe une autre fonction g telle que, pour tout $A \in \mathcal{F}$,

$$\mu(A) = \int_A g \, d\nu,$$

alors, $f = g$ ν-p.p. (ν-presque-partout). Par convention, si $(\Omega, \mathcal{F}) = (\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ et si la mesure ν n'est pas précisée (on dit que μ a une densité f) alors il est sous-entendu que ν est la mesure de Lebesgue k-dimensionnelle.

On dira que μ est **absolument continue** par rapport à ν, ce que l'on note $\mu \ll \nu$, si pour tout ensemble $A \in \mathcal{F}$ tel que $\nu(A) = 0$, nous avons $\mu(A) = 0$. Les mesures μ et ν sont **équivalentes**, $\mu \equiv \nu$ si nous avons simultanément $\mu \ll \nu$ et $\nu \ll \mu$.

Le lemme suivant montre que ces notions sont préservées par passage aux *mesures images*.

Lemme A.5.1
Supposons que $\mu \ll \nu$ et soit $T : \Omega \to \mathcal{T}$ une fonction mesurable de $(\Omega, \mathcal{F})$ dans $(\mathcal{T}, \mathcal{B}(\mathcal{T}))$. Notons μ_T et ν_T les mesures images de T définies à partir de μ et ν respectivement. Alors, $\mu_T \ll \nu_T$.

Démonstration. Pour tout $B \in \mathcal{B}(\mathcal{T})$,

$$\nu_T(B) = 0 \quad \Leftrightarrow \quad \nu(T^{-1}(B)) = 0 \quad \Leftrightarrow \quad \mu(T^{-1}(B)) = 0 \quad \Leftrightarrow \quad \mu_T(B) = 0.$$

D'où le résultat. ■

Supposons que $(\Omega, \mathcal{F}) = (\mathcal{X}, \mathcal{B}(\mathcal{X}))$, où $\mathcal{X}$ est l'espace $\mathbb{R}^k$ muni de la métrique induite par la distance euclidienne. Une mesure positive μ sur $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ telle que, pour tout ensemble borné $A \in \mathcal{B}(\mathcal{X})$, $\mu(A)$ est finie est appelée **mesure positive σ-finie**. Il est clair que si μ a une densité par rapport à ν, alors $\nu \ll \mu$. Ce résultat admet une réciproque que nous allons admettre.

Théorème A.5.2 (Théorème de Radon-Nikodym)
Soient μ et ν deux mesures σ-finies. Une condition nécessaire et suffisante pour que $\mu \ll \nu$ est que μ admet une densité f par rapport à ν.

La fonction f est aussi appelée la dérivée de Radon-Nikodym de la mesure μ par rapport à la mesure ν, et on la note usuellement $f = d\mu/d\nu$, ou encore $d\mu = f d\nu$. Toutes les propriétés des dérivées de Radon-Nikodym suggérées par cette écriture différentielle sont vérifiées. A titre d'exemples : si $d\mu_1 = f_1 d\nu$ et $d\mu_2 = f_2 d\nu$, alors $d(\mu_1 + \mu_2) = (f_1 + f_2) d\nu$; si $d\lambda = f d\mu$ et $d\mu = g d\nu$ alors, $d\lambda = f g d\nu$.

En théorie des distributions, on appelle **mesure de Radon** (nous allons voir pourquoi le terme *mesure* est appropriée) une forme linéaire positive u définie sur l'espace $C_0(\mathcal{X})$ des fonctions continues à support compact muni de la norme sup. La positivité signifie ici que si f est une fonction à valeurs positives alors $< u, f > \geq 0$ (on utilise ici la notation classique $< u, f >$ pour $u(f)$ quand u est une forme linéaire). On montre facilement qu'une telle forme linéaire vérifie la propriété de continuité suivante. Pour tout compact K, il existe C tel que pour tout $f \in C_0(\mathcal{X})$ à support dans K, $| < u, f > | \leq C \sup |f|$ ce qui fait de u une

distribution. Il est facile de voir qu'une mesure positive σ-finie μ définit une mesure de Radon à travers l'application

$$f \mapsto \int f \, d\mu$$

définie sur $C_0(\mathcal{X})$. Nous concluons ce paragraphe succint sur les mesures σ-finies par un résultat fondamental d'analyse qui met en relation théorie de la mesure et théorie des distributions en apportant une réciproque à ce résultat.

Théorème A.5.3 (Théorème de représentation de Riesz)
Pour toute mesure de Radon u, il existe une unique mesure positive σ-finie μ définie sur $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ telle que, pour tout $f \in C_0(\mathcal{X})$, $< u, f >= \int f \, d\mu$.

A.6 Moments d'ordre p, espaces $\mathcal{L}^p$ et L^p

Soit X une v.a. à valeurs réelle. Pour $p > 0$, on dit que X **admet un moment d'ordre** p si $|X|^p$ admet un moment d'ordre un c'est-àdire, $\mathbb{E}[|X|^p] < \infty$. Nous notons $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ ($\mathcal{L}^p$ pour faire court) l'espace des variables aléatoires définies sur $(\Omega, \mathcal{F})$ admettant un moment d'ordre p parapport à la mesure $\mathbb{P}$. Nous notons, pour $X \in \mathcal{L}^p$, $\|X\|_p = (\mathbb{E}[|X|^p])^{1/p}$. Cette définition s'étend au cas $p = \infty$ par la *borne essentielle* de X définie par :

$$\|X\|_\infty = \sup \{a; \mathbb{P}\{|X| > a\} > 0\} .$$

Les inégalités suivantes sont souvent utiles

Proposition A.6.1
Soit $1 \leq p \leq r \leq \infty$ et $Y \in \mathcal{L}^r$. Alors, $Y \in \mathcal{L}^p$ et $\|Y\|_p \leq \|Y\|_r$.

Cette inégalité est triviale dans le cas $r = \infty$ et, dans le cas $r < \infty$, découle directement de l'inégalité de Jensen en remarquant que $x \mapsto x^{r/p}$ est convexe sur $\mathbb{R}_+$ (voir Proposition A.4.3).

Proposition A.6.2
*Soit $p \geq 1$. Nous avons (**inégalité de Minkowski**) :*

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

*Soient $p, q \geq 1$ tels que $p^{-1} + q^{-1} = 1$. Nous avons (**inégalité de Hölder**) :*

$$\|XY\|_1 \leq \|X\|_p \|Y\|_q.$$

L'inégalité de Hölder pour $p = 2$ est aussi appelée **inégalité de Schwarz**. On en déduit que $\| \bullet \|_p$ est positive et vérifie l'inégalité triangulaire. On voit de blus que $\|\lambda \bullet \|_p = |\lambda| \| \bullet \|_p$ pour tout réel λ. Ce n'est toutefois pas une norme, car la relation $\|X\|_p = 0$ entraîne seulement que $X = 0$ $\mathbb{P}$-p.s. On dit que $\| \bullet \|_p$ est **une semi-norme**. Comme nous le verrons ci-dessus, il est possible (mais pas toujours utile ni pratique), de "quotienter" l'espace par la relation d'équivalence $X \equiv Y \iff X = Y$, $\mathbb{P}$-p.s. La proposition suivante permet de montrer que l'espace quotienté est un espace de Banach.

Proposition A.6.3
Soit $p \in [1, \infty)$. Soit (X_n) une suite de Cauchy dans $\mathcal{L}^p$, i.e., :

$$\lim_{k \to \infty} \sup_{r,s \geq k} \|X_r - X_s\|_p = 0.$$

Il existe une variable aléatoire $X \in \mathcal{L}^p$ telle que $X_r \to X$ dans $\mathcal{L}_p$, i.e. $\|X_r - X\|_p \to 0$. De plus, on peut extraire de X_n une sous-suite $Y_k = X_{n_k}$ qui converge vers X $\mathbb{P}$-p.s.

Démonstration. C'est un résultat classique d'analyse ; nous en donnons toutefois une démonstration de nature "probabiliste" afin d'illustrer les résultats et les techniques introduites précédemment. Soit $k_n \nearrow \infty$ une suite telle que :

$$\forall (r,s) \geq k_n, \|X_r - X_s\|_p \leq 2^{-n}.$$

Nous avons, par monotonicité des semi-normes $\| \bullet \|_p$:

$$\mathbb{E}\left[|X_{k_{n+1}} - X_{k_n}|\right] \leq \|X_{k_{n+1}} - X_{k_n}\|_p \leq 2^{-n},$$

ce qui implique que $\mathbb{E}\left[\sum_n |X_{k_{n+1}} - X_{k_n}|\right] < \infty$. Donc, la série de terme général $U_n = (X_{k_{n+1}} - X_{k_n})$ converge absolument $\mathbb{P}$-p.s. et donc $\lim_{n\to\infty} X_{k_n}$ existe p.s. Définissons, pour tout $\omega \in \Omega$:

$$X(\omega) := \limsup X_{k_n}(\omega).$$

X est une v.a. (en tant que limite supérieure d'une suite de v.a.) et $X_{k_n} \to X$ p.s. Soit $r \in \mathbb{N}$ et soit $n \in \mathbb{N}$ tel que $r \geq k_n$; pour tout $m \geq n$, on a :

$$\|X_r - X_{k_m}\|_p \leq 2^{-n},$$

et l'application du lemme de Fatou montre que :

$$\|X_r - X\|_p \leq \liminf_m \|X_r - X_{k_m}\|_p \leq 2^{-n}.$$

Cette relation montre que $(X_r - X) \in \mathcal{L}_p$ et donc que $X \in \mathcal{L}^p$; de plus, cette relation montre que $X_r \to X$ dans $\mathcal{L}_p$. ■

Le résultat précédent montre que $\mathcal{L}^p$ peut-être muni d'une structure d'espace vectoriel normé complet par passage au quotient. Deux variables aléatoires X et Y sont égales presque-sûrement, si $\mathbb{P}\{\omega : X(\omega) = Y(\omega)\} = 1$. L'égalité presque-sûre sur $(\Omega, \mathcal{F}, \mathbb{P})$ définit une relation d'équivalence sur l'ensemble des v.a. à valeurs dans $(E, \mathcal{E})$. Si X et Y sont deux éléments de la même classe d'équivalence, et si X admet un moment d'ordre p, alors $\mathbb{E}\left[|X|^p\right] = \mathbb{E}\left[|Y|^p\right]$. Lorsque l'on choisit un élément d'une classe d'équivalence on dit que l'on choisit une **version** de X. Dans la suite, nous utiliserons la même notation X pour la v.a., la classe d'équivalence de X (l'ensemble des v.a. égales à X p.s.) et n'importe quel autre élément de la classe d'équivalence de X (ou version de la classe de X).

On note $L^p(\Omega, \mathcal{F}, \mathbb{P})$ l'espace des classes d'équivalence des variables de $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$. La proposition A.6.3 montre que $L^p(\Omega, \mathcal{F}, \mathbb{P})$ muni de la norme $\| \bullet \|_p$ est un espace vectoriel normé complet, c'est-à-dire **un espace de Banach**.

A.7 Variance, covariance

Soit X une variable aléatoire admettant un moment d'ordre 2 ; alors X admet un moment d'ordre 1 (par monotonicité des semi-normes, $\mathcal{L}^1 \subset \mathcal{L}^2$). On pose alors :

$$\mathrm{var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2) = \mathbb{E}[X^2] - \mathbb{E}^2[X],$$

quantité que l'on appelle la *variance* de X. De même, lorsque $X, Y \in \mathcal{L}^2$, nous pouvons définir :

$$\mathrm{cov}(X,Y) := \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right],$$

quantité que l'on appelle la covariance de X et de Y. Les variables aléatoires sont dites "décorrélées", si le coefficient de covariance $\mathrm{cov}(X,Y) = 0$. Lorsque $X := (X_1, \cdots, X_d)^T$, $d \in \mathbb{N}$, la matrice de covariance $\Gamma(X)$ (ou matrice de variance / covariance) est définie comme la matrice $d \times d$:

$$\Gamma(X)_{i,j} = \mathrm{cov}(X_i, X_j)$$

Les éléments diagonaux sont égaux à la variance des variables X_i ; les éléments hors-diagonaux sont les coefficients de covariance. La matrice de covariance est une matrice symétrique ($\Gamma(X) = \Gamma(X)^T$) et semi-définie positive. En effet, pour tout d-uplets de nombre complexes $(a_1, a_2, \cdots, a_d)$, nous avons :

$$\mathbb{E}\left[\left(\sum_{i=1}^{d} a_i(X_i - \mathbb{E}[X_i])\right)^2\right] = \sum_{i,j} a_i a_j^* \Gamma(X)_{i,j} \geq 0$$

Notons que, pour tout vecteur a (déterministe) :

$$\Gamma(X + a) = \Gamma(X),$$

et que, pour M une matrice (déterministe) $p \times d$:

$$\Gamma(MX) = M\Gamma(X)M^T.$$

Nous munissons l'espace $\mathcal{L}^2$ du produit scalaire :

$$< X, Y >:= \mathbb{E}[XY]$$

Comme précédemment toutefois, ce produit scalaire n'induit pas une norme, mais une semi-norme (voir ci-dessus). Définissons L^2 l'espace quotient de $\mathcal{L}^2$ par la relation d'équivalence d'égalité p.s. Le produit scalaire défini ci-dessus s'étend directement à l'espace quotient, car pour toutes variables $\tilde{X}$ (resp. $\tilde{Y}$) de la classe de X (resp. Y), nous avons

$$< \tilde{X}, \tilde{Y} >=< X, Y > .$$

On vérifie aisément que L^2 muni de ce produit scalaire est un espace hilbertien. Cette propriété a un grand nombre de conséquences. Nous utiliserons en particulier cette propriété pour construire l'espérance conditionnelle.

A.8 Indépendance. Mesures produits

Soient A et B deux événements. On dit que A et B sont *indépendants* si :

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

Les propriétés élémentaires des probabilités montrent que les événements A et B^c, A^c et B, et A^c et B^c sont aussi indépendants. En effet :

$$\mathbb{P}(A^c \cap B) = \mathbb{P}(\Omega \cap B) - \mathbb{P}(A \cap B) = \mathbb{P}(B) - \mathbb{P}(A)\mathbb{P}(B) = (1 - \mathbb{P}(A))\mathbb{P}(B).$$

Les tribus $\mathcal{A} = \{\emptyset, A, A^c, \Omega\}$ et $\mathcal{B} = \{\emptyset, B, B^c, \Omega\}$ sont donc indépendantes, au sens de la définition suivante

Définition A.8.1 (Indépendance)**.** *Soit $(\mathcal{B}_i, i \in I)$ une famille de tribu. On dit que cette famille est* indépendante *si, pour tout sous-ensemble J fini de I :*

$$\mathbb{P}\left(\bigcap_{j\in J} B_j\right) = \prod_{j\in J} \mathbb{P}(B_j), \quad B_j \in \mathcal{B}_j$$

Le lemme technique suivant donne un critère plus "pratique" pour vérifier l'indépendance de tribus.

Lemme A.8.2
Soient $\mathcal{G}$ et $\mathcal{H}$ deux sous-tribus de $\mathcal{F}$ et soit $\mathcal{I}$ et $\mathcal{J}$ deux π-systèmes tels que $\mathcal{G} := \sigma(\mathcal{I})$ et $\mathcal{H} := \sigma(\mathcal{J})$. Alors, les tribus $\mathcal{G}$ et $\mathcal{H}$ sont indépendantes si et seulement si $\mathcal{I}$ et $\mathcal{J}$ sont indépendantes, i.e. *:*

$$\mathbb{P}(I \cap J) = \mathbb{P}(I)\mathbb{P}(J), \quad I \in \mathcal{I}, J \in \mathcal{J}.$$

Démonstration. Supposons que les familles $\mathcal{I}$ et $\mathcal{J}$ sont indépendantes. Pour $I \in \mathcal{I}$ donné, considérons les mesures :

$$H \to \mathbb{P}(I \cap H) \text{ et } H \to \mathbb{P}(I)\mathbb{P}(H).$$

Ces mesures sont définies $(\Omega, \mathcal{H})$ et coincident sur $\mathcal{J}$. Le théorème A.2.6 montre que ces deux mesures coïncident sur $\mathcal{H}$:

$$\mathbb{P}(I \cap H) = \mathbb{P}(I)\mathbb{P}(H), \quad I \in \mathcal{I}, H \in \mathcal{H}.$$

Pour H donné dans $\mathcal{H}$, les mesures :

$$G \to \mathbb{P}(G \cap H) \text{ et } G \to \mathbb{P}(G)\mathbb{P}(H)$$

sont définies sur $\mathcal{G}$ et coïncident sur $\mathcal{I}$. Par le théorème extension, elles coïncident sur $\mathcal{G}$, et donc $\mathbb{P}(G \cap H) = \mathbb{P}(G)\mathbb{P}(H)$, pour tout $G \in \mathcal{G}$ et $H \in \mathcal{H}$. ■

De façon générale, on a

Proposition A.8.3
Soient $(\mathcal{C}_i, i \in I)$ une famille de π-systèmes indépendants. Alors les tribus $(\sigma(\mathcal{C}_i), i \in I)$ sont indépendantes.

Il résulte immédiatement de la définition A.8.1 que si $\mathcal{B}'_i$ est une sous-tribu de $\mathcal{B}_i$, la famille $(\mathcal{B}'_i, i \in I)$ est une famille indépendante si $(\mathcal{B}_i, i \in I)$ l'est. Nous avons aussi

Proposition A.8.4
Si la famille $(\mathcal{B}_i, i \in I)$ est indépendante et si $(I_j, j \in J)$ est une partition de I, la famille $(\sigma(B_i, i \in I_j), j \in J)$ est indépendante.

De cette définition découle toutes les notions d'indépendance dont nous aurons besoin dans la suite. Si $(A_i, i \in I)$ est une famille d'événements, on dira que cette famille est indépendante si la famille $(\sigma(A_i), i \in I)$ l'est. Si $(X_i, i \in I)$ est une famille de v.a., on dira que cette famille est indépendante si la famille $(\sigma(X_i), i \in I)$ l'est. Si X est une v.a. et $\mathcal{G}$ une tribu, on dira que X et $\mathcal{G}$ sont indépendantes si les tribus $\sigma(X)$ et $\mathcal{G}$ sont indépendantes. Enfin, si $(X_i, i \in I)$ et $(Y_j, j \in J)$ sont indépendantes si les tribus $(\sigma(X_i), i \in I)$ et $(\sigma(Y_j), j \in J)$ le sont.

Exemple A.3:
Soient (X_1, X_2, X_3, X_4) quatre v.a. indépendantes. Alors, les couples (X_1, X_2) et (X_3, X_4) sont indépendants, puisque les tribus $\sigma(X_1, X_2)$ et $\sigma(X_3, X_4)$ le sont. Alors $Y_1 := f(X_1, X_2)$ et $Y_2 = g(X_3, X_4)$ (avec f, g boréliennes) sont indépendantes car $\sigma(Y_1) \subset \sigma(X_1, X_2)$ et $\sigma(Y_2) \subset \sigma(X_3, X_4)$.

Avant d'aller plus loin, rappelons quelques résultats sur les mesures produits (on se reportera avec profit au cours d'intégration). Soient $(E_1, \mathcal{B}_1, \nu_1)$ et $(E_2, \mathcal{B}_2, \nu_2)$ deux espaces mesurés et ν_1, ν_2 deux mesures σ-finies. Alors :

$$\mathcal{B}_1 \otimes \mathcal{B}_2 := \sigma(A_1 \times A_2, A_1 \in \mathcal{B}_1, A_2 \in \mathcal{B}_2)$$

est une tribu sur $E_1 \times E_2$ appelée *tribu produit* de $\mathcal{B}_1$ et de $\mathcal{B}_2$ et il existe une unique mesure, notée $\nu_1 \otimes \nu_2$ définie sur $\mathcal{B}_1 \otimes \mathcal{B}_2$ telle que :

$$\nu_1 \otimes \nu_2(A_1 \times A_2) = \nu_1(A_1)\nu_2(A_2), \ A_1 \in \mathcal{B}_1, A_2 \in \mathcal{B}_2.$$

Nous rappelons le théorème fondamental suivant.

Théorème A.8.5 (Théorème de Fubini)
Soit $f : E_1 \times E_2 \to \mathbb{R}$ une fonction mesurable positive (où on a muni $E_1 \times E_2$ de la tribu $\mathcal{B}_1 \otimes \mathcal{B}_2$. Alors

$$\begin{aligned} \int f \, d(\nu_1 \otimes \nu_2) &= \int \left(\int f(x_1, x_2) d\nu_1(x_1) \right) d\nu_2(x_2), \\ &= \int \left(\int f(x_1, x_2) d\nu_2(x_2) \right) d\nu_1(x_1) \end{aligned}$$

Il s'en suit que, pour toute fonction mesurable $f : E_1 \times E_2 \to \mathbb{R}$, f est $(\nu_1 \otimes \nu_2)$-intégrable si et seulement si

$$int \left(\int |f|(x_1, x_2) d\nu_1(x_1) \right) d\nu_2(x_2) < \infty$$

(ou dans l'ordre inverse) et, si c'est le cas,

$$\begin{aligned} \int f \, d(\nu_1 \otimes \nu_2) &= \int \left(\int f(x_1, x_2) d\nu_1(x_1) \right) d\nu_2(x_2), \\ &= \int \left(\int f(x_1, x_2) d\nu_2(x_2) \right) d\nu_1(x_1) \end{aligned}$$

Ces résultats s'étendent directement pour le produit de n espaces. Il résulte alors de ces rappels et du théorème de classe monotone que

Théorème A.8.6
Soient $(X_1, \cdots, X_n)$ des v.a. à valeurs dans $(E_i, \mathcal{E}_i)$, $i \in \{1, \cdots, n\}$. Il y a équivalence entre

1. *les v.a $X_1, \cdots, X_n$ sont indépendantes,*
2. *Pour tout $A_k \in \mathcal{E}_k$:*

$$\mathbb{P}[X_1 \in A_1, \cdots, X_n \in A_n] = \prod_1^n \mathbb{P}[X_k \in A_k]$$

3. *Pour tout $A_k \in \mathcal{C}_k$, avec $\mathcal{C}_k$ π-système tel que $\sigma(\mathcal{C}_k) = \mathcal{E}_k$:*

$$\mathbb{P}[X_1 \in A_1, \cdots, X_n \in A_n] = \prod_1^n \mathbb{P}[X_k \in A_k]$$

4. *La loi du vecteur aléatoire $(X_1, \ldots, X_n)$, notée $\mathbb{P}_{(X_1,\cdots,X_n)}$ est égale au produit des lois des v.a X_k :*

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

5. *Pour toutes fonctions f_k boréliennes positives (resp. bornées, resp. $f_k \in \mathcal{L}^1(E_k, \mathcal{E}_k, \mathbb{P}_k)$) :*

$$\mathbb{E}\left[f_1(X_1) \ldots f_n(X_n)\right] = \prod_1^n \mathbb{E}\left[f_k(X_k)\right].$$

Exemple A.4:
Soient X, Y deux v.a. Alors, vu que $\sigma([a, b[, a < b \in \mathbb{R}) = \mathcal{B}(\mathbb{R})$, il résulte du théorème précédent que X et Y sont indépendantes si et seulement si :

$$\mathbb{P}(a \leq X < b, c \leq Y < d) = \mathbb{P}(a \leq X < b)\mathbb{P}(c \leq Y < d),$$

pour tout a, b, c, d. Dans ce cas, si $\mathbb{E}\left[|X|\right] < \infty$, $\mathbb{E}\left[|Y|\right] < \infty$, on a $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, résultat que l'on utilise sans cesse en probabilité.

A.9 Fonction caractéristique

Soit X une variable aléatoire à valeurs dans $\mathbb{R}^d$. L'application $\Phi : \mathbb{R}^d \to \mathbb{C}$ définie par :

$$\Phi(\lambda) = \mathbb{E}[\exp(\mathrm{i}\lambda^T X)]\ ,$$

est appellée la *fonction caractéristique* de X. Nous donnons ci-dessous quelques propriétés élémentaires de la fonction caractéristique

(i) $\Phi(0) = 1$ et $|\Phi(\lambda)| \leq 1$.

(ii) La fonction caractéristique est continue sur $\mathbb{R}^d$. Cette propriété est une conséquence immédiate de la continuité de l'application $\lambda \longrightarrow \exp(\mathrm{i}\lambda^T X)$ et du théorème de convergence dominé.

(iii) Lorsque la loi de X admet une densité g par rapport à la mesure de Lebesgue, alors Φ est la transformée de Fourier de g (au sens usuel). Le théorème de Rieman-Lebesgue implique alors que $\Phi(\lambda)$ tend vers 0 lorsque $\lambda \to \infty$.

La propriété ii se généralise sous la forme du résultat suivant.

Proposition A.9.1
Soit X une variable aléatoire à valeurs dans $\mathbb{R}$ et $k \in \mathbb{N}$. Si $\mathbb{E}[|X|^k] < \infty$, alors la fonction caractéristique Φ de X est k fois continûment dérivable sur $\mathbb{R}$ et admet en $\lambda = 0$ le développement de Taylor :

$$\Phi(\lambda) = \sum_{j=0}^{k} \frac{\mathbb{E}[X^j]}{j!} (\mathrm{i}\lambda)^j + o(\lambda^k)\ .$$

Comme son nom l'indique, la fonction caractéristique "caractérise" la loi, dans le sens

Proposition A.9.2
Deux variables aléatoires à valeurs dans $\mathbb{R}^d$ ont même loi si et seulement si elles ont même fonction caractéristique.

Le théorème précédent admet le corollaire suivant,

Proposition A.9.3
Soient X et Y deux variables aléatoires à valeurs dans $\mathbb{R}^{d_1}$ et $\mathbb{R}^{d_2}$. Ces deux variables sont indépendantes si et seulement si pour tout $\lambda_1 \in \mathbb{R}^{d_1}$ et $\lambda_2 \mathbb{R}^{d_2}$

$$\mathbb{E}[\exp(\mathrm{i}[\lambda_1^T \lambda_2^T][X^T Y^T]^T)] = \mathbb{E}[\exp(\mathrm{i}\lambda_1^T X)]\ \mathbb{E}[\exp(\mathrm{i}\lambda_2^T Y)]\ .$$

De plus dans ce cas, si $d_1 = d_2$, pour tout $\lambda \in \mathbb{R}^{d_1}$

$$\mathbb{E}[\exp(\mathrm{i}[\lambda^T (X+Y))] = \mathbb{E}[\exp(\mathrm{i}\lambda^T X)]\ \mathbb{E}[\exp(\mathrm{i}\lambda^T Y)]\ .$$

A.10 Fonction génératrice des moments

La **fonction génératrice** est l'extension de la fonction caractéristique aux *valeurs complexes* de λ. Soit X une v.a. réelle, sa fonction génératrice $M_X(z)$ est définie par :

$$M_X(z) = \mathbb{E}[z^X],$$

pour tout $z \in \mathbb{C}$ où cette quantité est bien définie, ce qui est au moins le cas sur le cercle complexe unité. Dans le cas où X est à valeur dans $\mathbb{N}$, on montre facilement que $M_X(z)$ est un série entière de rayon de convergence au moins égal à 1 entièrement caractérisée par la loi de X. De plus, dans ce cas, on a, pour X et Y indépendantes,

$$M_{X+Y} = M_X\, M_Y.$$

Cette propriété s'avère souvent utile pour caractériser la loi d'une somme de v.a. entières indépendantes.

A.11 Espérance conditionnelle

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité. Pour tout $A, B \in \mathcal{F}$, on appelle *probabilité conditionnelle* de A sachant B la quantité :

$$\begin{cases} \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} & \text{si } \mathbb{P}(B) > 0 \\ \mathbb{P}(A|B) = 0 & \text{sinon.} \end{cases}$$

On remarque alors que pour tout $B \in \mathcal{F}$ tel que $\mathbb{P}(B) > 0$, l'application $A \mapsto \mathbb{P}(A|B)$ définit une probabilité sur $(\Omega, \mathcal{F})$. Cette probabilité s'appelle la *loi conditionnelle sachant l'événement* B et l'espérance d'une v.a. X par rapport à cette probabilité est l' *espérance conditionnelle* de X *sachant l'événement* B, notée :

$$\mathbb{E}[X|B] = \frac{1}{\mathbb{P}(B)} \int_B X(\omega)\, d\mathbb{P}(\omega).$$

L'espérance conditionnelle $\mathbb{E}[X|B]$ représente l'espérance de la variable aléatoire X *sachant* que l'événement B est réalisé.

Exemple A.5:
Soit X une variable aléatoire à valeurs dans l'ensemble des entiers naturels $\mathbb{N}$. La loi de X est spécifiée par la donnée des probabilités $p_i = \mathbb{P}(X = i)$, pour $i \in \mathbb{N}$. La moyenne de X est donnée par $\mathbb{E}[X] = \sum_{i\in\mathbb{N}} ip_i$. Considérons l'événement $B = \{X \geq i_0\}$. Nous avons $\mathbb{P}(B) = \sum_{i\geq i_0} p_i$ que nous supposerons non nul par hypothèse. L'espérance conditionnelle de X sachant B est donnée par :

$$\mathbb{E}[X|B] = \frac{\sum_{i\geq i_0} ip_i}{\sum_{i\geq i_0} p_i} \tag{A.7}$$

qui correspond à la moyenne de X conditionnellement à l'événement $B = \{X \geq i_0\}$.

Exemple A.6:
Soient X et T deux variables aléatoires définies sur $(\Omega, \mathcal{F}, \mathbb{P})$. Supposons X à valeurs dans $\mathcal{X}$ et T à valeurs dans un ensemble discret $\{t_i : i \in \mathbb{N}\}$. Alors, pour tout $i \in \mathbb{N}$ et toute statistique $\phi(X)$ positive ou intégrable, l'espérance conditionnelle de $\phi(X)$ sachant $\{T = t_i\}$ s'écrit, quand $\mathbb{P}(T = t_i) > 0$,

$$\mathbb{E}[\phi(X)|T = t_i] = \frac{1}{\mathbb{P}(T = t_i)} \int_{T=t_i} \phi(X)\, d\mathbb{P} = \int \phi(x)\, \mathbb{P}_{X|Y}(dx|t_i),$$

où nous avons introduit la notation $\mathbb{P}_{X|Y}$ pour la mesure de probabilité définie sur $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ définie par $A \mapsto \mathbb{P}(X \in A|T = t_i)$.

Dans l'exemple A.6, on a défini une famille de probabilité paramétrée par t_i, où t_i décrit les valeurs prises par la variables aléatoire T. On peut se demander si cette définition a un équivalent pour des variables aléatoires plus générales qu'une variable discrète. Cette généralisation se fait en deux temps. On commence par définir l'espérance conditionnelle de X sachant $\mathcal{B}$ pour X v.a. fixée et $\mathcal{B}$ une sous-tribu de $\mathcal{F}$. Puis, dans le cas où $\mathcal{B}$ est la tribu engendrée par une v.a. T, cette notion d'espérance conditionnelle permettra de généraliser l'exemple A.6 au cas où T n'est pas à valeurs discrètes.

Théorème A.11.1
Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et $\mathcal{B} \subset \mathcal{F}$ une sous-tribu de $\mathcal{F}$. Soit X une v.a. intégrable (resp. positive). Il existe une v.a. Y intégrable (resp. positive) $\mathcal{B}$-mesurable, telle que :

$$\forall B \in \mathcal{B}, \int_B X\, d\mathbb{P} = \int_B Y\, d\mathbb{P}. \tag{A.8}$$

Cette variable est unique à une $\mathbb{P}$-équivalence près.

Le cadre donné par le théorème A.11.1 permet la définition suivante.

Définition A.11.2. *Sous les hypothèses et les conclusions du théorème A.11.1, on appelle Y (en gardant à l'esprit que cette v.a. est définie à une équivalence près) l'*espérance conditionnelle *de X* sachant la tribu *$\mathcal{B}$, et on la note $\mathbb{E}[X \,|\, \mathcal{B}]$.*

Il faut faire attention que cette notion n'est pas une généralisation de (A.7) car dans ce dernier cas l'espérance conditionnelle est un nombre et dans le cas de la définition A.11.2, c'est une variable aléatoire $\mathcal{B}$-mesurable. Nous allons tout d'abord démontrer une version plus restrictive (et plus intuitive) du théorème A.11.1 en supposant que X est de carré intégrable. Nous verrons ensuite comment cette hypothèse peut être élargie. L'avantage de cette

hypothèse est de pouvoir utiliser la structure hilbertienne de l'espace L^2. En effet, l'espace $L^2 := L^2(\Omega, \mathcal{F}, \mathbb{P})$, muni du produit scalaire $< X, Y >:= \mathbb{E}[XY]$ est un espace hilbertien. Soit $\mathcal{B}$ une sous-tribu de $\mathcal{F}$ et définissons :

$$\mathcal{H}^{\mathcal{B}} := \left\{ Z \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P}), Z \quad \text{a un représentant } \mathcal{B} - \text{mesurable} \right\}.$$

Théorème A.11.3
Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et $\mathcal{B} \subset \mathcal{F}$ une sous-tribu de $\mathcal{F}$. Soit X une v.a., $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$. Il existe une unique (à une égalité presque sure près) variable aléatoire $Y \in \mathcal{H}^{\mathcal{B}}$ telle que :

$$\mathbb{E}\left[(X-Y)^2\right] = \inf_{Z \in \mathcal{H}^{\mathcal{B}}} \mathbb{E}\left[(X-Z)^2\right].$$

De façon équivalente, Y vérifie, pour toute v.a. $Z \in \mathcal{H}^{\mathcal{B}}$, $\mathbb{E}[XZ] = \mathbb{E}[YZ]$.

Démonstration. On remarque que les espérances apparaissant dans ce théorème ne dépendent pas des versions de Y ou Z. On peut donc raisonner sur les classes d'équivalences de ces variables. Notons $H^{\mathcal{B}}$ le quotient de $\mathcal{H}^{\mathcal{B}}$ par la relation d'équivalence d'égalité presque sure. On obtient $H = L^2(\Omega, \mathcal{B}, \mathbb{P})$, qui est un sous-espace vectoriel de $L^2(\Omega, \mathcal{F}, \mathbb{P})$ et il est fermé d'après la proposition A.6.3. Dans le cadre hilbertien, pour tout sous-espace fermé H de L^2, l'élément Y de H qui atteint $\inf_{Z \in H} \mathbb{E}\left[(X-Z)^2\right]$ est appelée la *projection* de X sur H. Montrons (résultat uniquement dû à la structure hilbertienne de l'espace L^2) que cette projection est bien définie de manière unique et qu'elle est caractérisée par la condition

(i) $\mathbb{E}[XZ] = \mathbb{E}[YZ]$ pour tout $Z \in H$.

Supposons que $Y \in H$ atteint $\inf_{Z \in H}$
$esp\left[(X-Z)^2\right]$ et montrons qu'il vérifie nécessairement (i). Pour tout $Z \in H$ et tout $t \in \mathbb{R}$, on a

$$0 \leq \mathbb{E}\left[(X-(Y+tZ))^2\right] - \mathbb{E}\left[(X-Y)^2\right] = t^2\mathbb{E}[Z^2] - 2t\mathbb{E}\left[(X-Y)Z\right].$$

Ceci n'est possible que si (i) est vérifié.

On remarque maintenant que, pour tout $Y, Z \in H$

$$\mathbb{E}\left[(X-Z)^2\right] = \mathbb{E}\left[(X-Y)^2\right] + \mathbb{E}\left[(Y-Z)^2\right] + 2\mathbb{E}\left[(X-Y)(Y-Z)\right].$$

Comme $Y - Z \in H$, on trouve donc que si Y vérifie la condition (i), alors il minimise $\mathbb{E}\left[(X-Z)^2\right]$ sur $Z \in H$ et tout autre élément différent de Y au sens L^2 ne minimise pas cette erreur. On obtient donc l'équivalence des deux deux conditions et l'unicité.

Il nous reste à montrer l'existence. Soit (Z_n) une suite de v.a. de H telle que $\mathbb{E}\left[(X-Z_n)^2\right] - m$ tendent vers zero, où $m = \inf_{Z \in F} \mathbb{E}\left[(X-Z)^2\right]$. On a, pour tout n, m,

$$\mathbb{E}\left[(X-Z_n-(X-Z_m))^2\right] + \mathbb{E}\left[(X-Z_n+(X-Z_m))^2\right] = 2\mathbb{E}\left[(X-Z_n)^2\right] + 2\mathbb{E}\left[(X-Z_m)^2\right]$$

D'où

$$\mathbb{E}\left[(Z_n-Z_m)^2\right] = 2(\mathbb{E}\left[(X-Z_n)^2\right] - m) + 2(\mathbb{E}\left[(X-Z_m)^2\right] - m) - 4(\mathbb{E}\left[\left(X - \frac{Z_n+Z_m}{2}\right)^2\right] - m).$$

Comme $(Z_n - Z_m)/2 \in H$, on a $\mathbb{E}\left[(X-(Z_n-Z_m)/2)^2\right] \geq m$. On obtient donc que (Z_n) est de Cauchy. Comme H est fermé et sous-ensemble d'un espace de Banach, il est complet et Z_n converge donc dans H, ce qui achève la preuve de l'existence. ■

Nous aurons besoin du lemme élémentaire d'unicité suivant

Lemme A.11.4
Soient X et Y deux v.a. $\mathcal{B}$-mesurables toutes deux positives ou toutes deux intégrables vérifiant :

$$\forall B \in \mathcal{B},\ \int_B X\, d\mathbb{P} \geq \int_B Y\, d\mathbb{P} \quad (\textit{respectivement} =)$$

Alors, $X \geq Y$ $\mathbb{P}$-p.s. (respectivement $X = Y$, $\mathbb{P}$-p.s.).

Démonstration. pour $a < b$, définissons $F_{a,b} := \{X \leq a < b \leq Y\} \in \mathcal{B}$. Puisque $\{X < Y\} = \bigcup_{a,b\in\mathbb{Q}} F_{a,b}$, il suffit de prouver que, pour tout $a, b \in \mathbb{Q}$, $\mathbb{P}(F_{a,b}) = 0$. Mais si $\mathbb{P}(F_{a,b}) > 0$ nous avons :

$$\int_{F_{a,b}} X\, d\mathbb{P} \leq a\mathbb{P}(F_{a,b}) < b\mathbb{P}(F_{a,b}) \leq \int_{F_{a,b}} Y\, d\mathbb{P}$$

et nous aboutissons à une contradiction. ■

Démonstration. Théorème A.11.1 L'unicité découle du lemme A.11.4. Montrons l'existence. On suppose tout d'abord que $X \geq 0$. Pour $n \in \mathbb{N}$, définissons $X_n = \min(X, n)$. $X_n \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$, et il existe donc une v.a. $Y_n \geq 0$, $\mathcal{B}$-mesurable, unique à une équivalence près, telle que :

$$\forall B \in \mathcal{B},\quad \int_B X_n\, d\mathbb{P} = \int_B Y_n\, d\mathbb{P}.$$

Par application de A.11.4, Y_n est $\mathbb{P}$-p.s. une suite positive et croissante. En effet, pour tout $B \in \mathcal{B}$:

$$\int_B Y_{n+1}\, d\mathbb{P} = \int_B X_{n+1}\, d\mathbb{P} \geq \int_B X_n\, d\mathbb{P} = \int_B Y_n\, d\mathbb{P}.$$

Définissons $Y = \lim \nearrow Y_n$. Y est $\mathcal{B}$-mesurable et, par application du théorème de convergence monotone, pour tout $B \in \mathcal{B}$, nous avons :

$$\int_B Y\, d\mathbb{P} = \lim \nearrow \int_B Y_n\, d\mathbb{P} = \lim \nearrow \int_B X_n\, d\mathbb{P} = \int_B X\, d\mathbb{P}.$$

Notons que si X est intégrable, alors Y l'est aussi (prendre $B = \Omega$). Pour étendre le résultat au cas intégrable, nous allons prouver que, pour X, Y deux v.a. positives intégrables, et pour $a, b \in \mathbb{R}$, nous avons (linéarité de l'espérance conditionnelle) :

$$\mathbb{E}[aX + bY|\mathcal{F}] = a\mathbb{E}[X|\mathcal{F}] + b\mathbb{E}[Y|\mathcal{F}].$$

Il suffit en effet de remarquer que, pour tout $B \in \mathcal{B}$

$$\begin{aligned}\int_B \mathbb{E}[aX + bY|\mathcal{F}]\, d\mathbb{P} &= \int_B (aX + bY)\, d\mathbb{P} = a\int_B X\, d\mathbb{P} + b\int Y\, d\mathbb{P} \\ &= a\int_B \mathbb{E}[X|\mathcal{B}]\, d\mathbb{P} + b\int_B \mathbb{E}[Y|\mathcal{B}]\, d\mathbb{P} = \int_B (a\mathbb{E}[X|\mathcal{B}] + b\mathbb{E}[Y|\mathcal{B}])\, d\mathbb{P}\end{aligned}$$

et on conclut en utilisant le lemme A.11.4. Pour $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, nous posons $X = X^+ - X^-$, et nous concluons en utilisant l'existence de l'espérance conditionnelle pour les v.a. positives et la linéarité de l'espérance conditionnelle. ■

Les propriétés suivantes découlent directement de la définition de l'espérance conditionnelle.

Proposition A.11.5 *1. Pour $X, Y \geq 0$ et $a, b \geq 0$ (ou X, Y intégrables et $a, b \in \mathbb{R}$), $\mathbb{E}[aX + bY|\mathcal{B}] = a\mathbb{E}[X|\mathcal{B}] + b\mathbb{E}[Y|\mathcal{B}]$.*

2. Pour $X, Y \geq 0$ (ou X, Y intégrables), l'inégalité $X \leq Y$ p.s. implique $\mathbb{E}[X|\mathcal{B}] \leq \mathbb{E}[Y|\mathcal{B}]$ p.s.

3. Soit $X \geq 0$. Alors $Y = \mathbb{E}[X|\mathcal{B}]$ vérifie, pour toute v.a. Z positive $\mathcal{B}$-mesurable, $\mathbb{E}[XZ] = \mathbb{E}[YZ]$.

4. Soit X intégrable. Alors $Y = \mathbb{E}[X|\mathcal{B}]$ vérifie, pour toute v.a. Z bornée $\mathcal{B}$-mesurable, $\mathbb{E}[XZ] = \mathbb{E}[YZ]$.

Citons quelques propriétés importantes de l'espérance qui s'étendent à l'espérance conditionnelle :

Proposition A.11.6 *1. ("Convergence monotone conditionnelle") Soit $(X_n)_{n\geq 0}$ une suite de v.a. telles que $0 \leq X_n \nearrow X$; alors $\mathbb{E}[X_n|\mathcal{G}] \nearrow \mathbb{E}[X|\mathcal{G}]$.*

2. ("Lemme de Fatou conditionnel") Soit $(X_n)_{n\geq 0}$ une suite de v.a. positives ; alors $\mathbb{E}[\liminf X_n|\mathcal{G}] \leq \liminf E[X_n|\mathcal{G}]$.

3. ("Convergence dominée conditionnelle") Soit $(X_n)_{n\geq 0}$ une suite de v.a. telle que $|X_n| \leq V$ $\mathbb{P}$-p.s., avec $\mathbb{E}[V] < \infty$ et $X_n \to X$ $\mathbb{P}$-p.s. Alors, $\mathbb{E}[X_n|\mathcal{G}] \to \mathbb{E}[X|\mathcal{G}]$ $\mathbb{P}$-p.s.

4. ("Inégalité de Jensen") Soit $c : \mathbb{R} \to \mathbb{R}$ convexe telle que $\mathbb{E}[|c(X)|] < \infty$. Alors, $\mathbb{E}[c(X)|\mathcal{G}] \leq c(\mathbb{E}[X|\mathcal{G}])$.

5. ("Contraction des normes") Pour $p \geq 1$, $\|\mathbb{E}[X|\mathcal{G}]\|_p \leq \|X\|_p$, en définissant $\|Y\|_p := (\mathbb{E}[|Y|^p])^{1/p}$.

Nous avons rassemblé dans la proposition suivante quelques propriétés essentielles de l'espérance conditionnelle, que nous utiliserons dans la suite.

Proposition A.11.7
Soit X une v.a. réelle.

1. Si $X \geq 0$ (ou X intégrable) et si $\mathcal{G}$ est la tribu grossière : $\mathcal{G} = \{\Omega, \emptyset\}$, alors $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X]$.

2. Si $X \geq 0$ (ou X intégrable) et $\mathcal{G} \subset \mathcal{B}$ deux sous-tribus de $\mathcal{F}$, alors

$$\mathbb{E}[\mathbb{E}[X|\mathcal{B}]|\mathcal{G}] = \mathbb{E}[X|\mathcal{G}].$$

3. Si $X \geq 0$ (ou X intégrable) est indépendant de $\mathcal{B}$ alors $\mathbb{E}[X|\mathcal{B}] = \mathbb{E}[X]$.

4. Si X est $\mathcal{B}$-mesurable et $Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ sont telles que $XY \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, alors $\mathbb{E}[XY|\mathcal{B}] = X\mathbb{E}[Y|\mathcal{B}]$.

Démonstration. Les fonctions mesurables par rapport à la tribu grossière sont les fonctions constantes. Donc, $\mathcal{G}$ étant la tribu grossière, $\mathbb{E}[X|\mathcal{G}] = c$. Par définition de l'espérance conditionnelle, nous avons :

$$\int_\Omega \mathbb{E}[X|\mathcal{G}]\, d\mathbb{P} = c = \int_\Omega X\, d\mathbb{P} = \mathbb{E}[X],$$

ce qui prouve la relation (1). Prouvons maintenant (2). Soit Z une v.a. $\mathcal{G}$-mesurable bornée. Notons que Z est aussi $\mathcal{B}$-mesurable. Par définition de l'espérance conditionnelle :

$$\mathbb{E}\left[\ \mathbb{E}[\mathbb{E}[X|\mathcal{B}]|\mathcal{G}]Z\ \right] = \mathbb{E}\left[\ \mathbb{E}[X|\mathcal{B}]Z\ \right] = \mathbb{E}\left[XZ\right] = \mathbb{E}\left[\mathbb{E}[X|\mathcal{G}]Z\right].$$

Donc, pour toute v.a. Z $\mathcal{B}$-mesurable bornée, $\mathbb{E}\left[\mathbb{E}[\mathbb{E}[X|\mathcal{B}]|\mathcal{G}]Z\right] = \mathbb{E}\left[\mathbb{E}[X|\mathcal{G}]Z\right]$, ce qui prouve la relation (2). Soit maintenant X une v.a. indépendante de $\mathcal{B}$. Alors, pour toute v.a. Z $\mathcal{B}$-mesurable bornée :

$$\mathbb{E}\left[\mathbb{E}[X|\mathcal{B}]Z\right] = \mathbb{E}\left[XZ\right] = \mathbb{E}\left[X\right]\mathbb{E}\left[Z\right] = \mathbb{E}\left[\mathbb{E}[X]Z\right],$$

ce qui prouve la relation (3). Considérons finalement la relation (4). Remarquons que $\mathbb{E}[Y|\mathcal{B}]X$ est $\mathcal{B}$-mesurable. Pour Z v.a. bornée $\mathcal{B}$-mesurable, on a, si on suppose X borné :

$$\mathbb{E}\left[\mathbb{E}[XY|\mathcal{B}]Z\right] = \mathbb{E}\{YXZ\} = \mathbb{E}\left[\mathbb{E}[Y|\mathcal{B}]XZ\right].$$

Ceci prouve la relation (4) pour X borné. Le cas général se prouve en utilisant la convergence dominée conditionnelle en posant $U_n = XY\mathbb{1}(|X| \leq n)$. ■

Nous introduisons maintenant la définition suivante correspondant au cas où la tribu $\mathcal{B}$ est engendrée par une v.a. (voir définition A.3.2).

Définition A.11.8. *Soit Y une v.a. définie sur $(\Omega, \mathcal{F}, \mathbb{P})$ et X une v.a. réelle intégrable ou positive définie sur le même espace. On appelle* espérance conditionnelle *de X* sachant *Y, et on note $\mathbb{E}[X\,|\,Y]$ la v.a. (définie à une équivalence près) $\mathbb{E}[X\,|\,\sigma(Y)]$.*

On sait d'après le théorème A.3.3 que $\mathbb{E}[X\,|\,Y]$ s'écrit $\phi(Y)$. Supposons que Y est à valeur dans $\mathcal{Y}$. Calculer $\mathbb{E}[X\,|\,Y]$ revient donc dans ce cas à trouver une fonction mesurable $\phi : \mathcal{Y} \to \mathbb{R}$ telle que $\mathbb{E}[X\,|\,Y] = \phi(Y)$. Il est courant de trouver la notation $\mathbb{E}[X\,|\,Y = y]$ pour $\phi(y)$, notation que l'on utilisera dans cet ouvrage. Il faut cependant faire attention de ne pas la confondre avec $\mathbb{E}[X\,|\,B]$, où B est l'événement $\{Y = y\}$, qui ne coïncide pas nécessairement avec $\phi(y)$ sauf si Y est à valeurs discrètes.

En pratique, la fonction $y \mapsto \mathbb{E}[X\,|\,Y = y]$ peut se calculer par des techniques de changement de variables, en faisant apparaître une variable Z indépendante de Y telle que $X = g(Y, Z)$ et en utilisant le lemme suivant

Lemme A.11.9
Soient Y et Z deux v.a. indépendantes. On supposera Y à valeurs dans $\mathcal{Y}$ et Z à valeurs dans $\mathcal{Z}$. Soit f une fonction mesurable de $\mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ telle que $\mathbb{E}|f(Y, Z)| < \infty$. Soit $\phi : \mathcal{Y} \to \mathbb{R}$ la fonction définie par

$$\phi(y) = \mathbb{E}[f(y, Z)]$$

Alors $\mathbb{E}[f(Y, Z)|Y] = \phi(Y)$.

Démonstration. Pour toute fonction mesurable bornée $\psi : \mathcal{Y} \to \mathbb{R}$, on a

$$\mathbb{E}[\phi(Y)\psi(Y)] = \int \left(\int f(Y(\omega_2), Z(\omega_1))\, d\mathbb{P}(\omega_1) \right) \psi(Y(\omega_2))\, d\mathbb{P}(\omega_2).$$

Par Fubini et par indépendance de Y et Z, cette dernière intégrale est précisément $\mathbb{E}[f(Y, Z)\psi(Y)]$, ce qui donne le résultat. ■

Néanmoins, dans la majeur partie des cas, une formule explicite et simple est donnée par un calcul de *densité conditionnelle* que nous introduisons maintenant. Soit (X, Y) un couple de v.a. à valeurs dans $\mathcal{X} = \mathbb{R}^k$ et $\mathcal{Y} = \mathbb{R}^l$ et définies sur le même espace de probabilité

$(\Omega, \mathcal{F}, \mathbb{P})$. Supposons que la loi de (X, Y) admet une densité $(x, y) \mapsto f(x, y)$ par rapport à une mesure dominante produit $\mu \otimes \nu$:

$$\mathbb{P}\{(X, Y) \in A \times B\} = \int_{A \times B} f(x, y)\, d\mu(x)\, d\nu(y), \quad A \in \mathcal{B}(\mathcal{X}),\, B \in \mathcal{B}(\mathcal{Y}).$$

Ceci implique que la loi de Y admet pour densité par rapport à la mesure dominante ν la fonction $y \mapsto f_Y(y)$ définie pour ν-presque tout y par

$$f_Y(y) = \int_{\mathcal{X}} f(x, y)\, d\mu(x).$$

On définit alors la *densité conditionnelle de X sachant $Y = y$* comme la fonction $x \mapsto f_{X|Y}(x|y)$ définie par

$$f_{X|Y}(x \,|\, y) := \frac{f(x, y)}{f_Y(y)}, \quad \text{pour tout } y \text{ tel que } f_Y(y) > 0$$

et prolongée arbitrairement si $f_Y(y) = 0$. Le résultat suivant s'applique alors.

Proposition A.11.10
Pour toute statistique $\psi(X)$ intégrable ou positive,

$$\mathbb{E}[\psi(X) \,|\, Y] = \int \psi(x)\, f_{X|Y}(x \,|\, Y)\, d\mu(x).$$

Démonstration. La preuve, relativement élémentaire, est laissée à titre d'exercice dans deux cas particuliers qui se généralisent très facilement : X et Y sont à valeur discrètes et $\mu \otimes \nu$ est la mesure de Lebesgue bi-dimensionnelle. ■

Nous concluons cette partie en introduisant la notion de *loi conditionnelle*.

Définition A.11.11. *Soient X et Y deux v.a. définies sur le même espace de régularité $(\Omega, \mathcal{F}, \mathbb{P})$. Supposons que X est à valeurs dans $\mathcal{X} = \mathbb{R}^k$ et Y à valeur dans $\mathcal{Y}$. On appelle* loi conditionnelle de X sachant Y *la fonction $\mathbb{P}_{X|Y} : \mathcal{B}(\mathcal{X}) \times \mathcal{Y} \to [0, 1]$ définie à une équivalence près par*

$$\text{pour tout } A \in \mathcal{B}(\mathcal{X}),\ \mathbb{P}_{X|Y}(A, Y) = \mathbb{E}[\mathbb{1}(X \in A) \,|\, Y] \quad p.s.$$

Dans cette définition, la mention *à une équivalence près* doit être comprise dans le sens suivant : pour tout $A \in \mathcal{B}(\mathcal{X})$, $\mathbb{P}_{X|Y}(A, \cdot)$ est une fonction mesurable définie sur $\mathcal{Y}$ définie à une $\mathbb{P}_Y$-équivalence près. On admet cependant que l'on peut choisir pour tout A une version de $\mathbb{P}_{X|Y}(A, \cdot)$ de telle sorte que, pour tout $y \in \mathcal{Y}$, $\mathbb{P}_{X|Y}(\cdot, y)$ est une loi de probabilité. Une telle version de $\mathbb{P}_{X|Y}$ est dite *version régulière* de la loi conditionnelle de X sachant Y. Elle définit alors une *probabilité de transition*, définie comme suit.

Définition A.11.12. *Soient $\mathcal{X}$ et $\mathcal{Y}$ deux espaces métriques que l'on munit de leurs boréliens. On dit que l'application $Q : \mathcal{Y} \times \mathcal{B}(\mathcal{X}) \to [0, 1]$ est une* probabilité de transition *(parfois aussi appelée* Noyau de probabilité*) si elle vérifie*

(1) pour tout $A \in \mathcal{B}(\mathcal{X})$, $Q(\cdot, A)$ est une fonction mesurable,

(2) pour tout $y \in \mathcal{Y}$, $Q(y, \cdot)$ est une loi de probabilité,

Il suit de ces définitions que, si $\mathbb{P}_{X|Y}$ est une version régulière de la loi conditionnelle de X sachant Y, alors pour toute statistique $f(X)$ positive ou intégrable, on a alors,

$$\mathbb{E}[f(X)|Y] = \int_{\mathcal{X}} f(x)\, \mathbb{P}_{X|Y}(dx, Y) \quad \text{p.s.}$$

Exemple A.7:
Si (X, Y) un couple de v.a. réelles définies sur $(\Omega, \mathcal{F}, \mathbb{P})$ tel que la loi de (X, Y) admet une densité f par rapport à une mesure dominante produit $\mu \otimes \nu$, nous avons vu que l'on pouvait définir la densité $f_Y(y)$ de Y par rapport à ν et la densité conditionnelle de X sachant Y notée $f_{X|Y}(x|y)$ définie pour tout $x \in \mathbb{R}$ et tout y tel que $f_Y(y) > 0$. On vérifie aisément que, pour tout $y \in \mathbb{R}$ tel que $f(y) > 0$, $f_{X|Y}(\cdot \,|\, y)\, d\mu(\cdot)$ est une mesure de probabilité. Si $f_Y(y) = 0$ on peut choisir n'importe quelle densité conditionnelle $f_{X|Y}(\cdot|y)$ qui ferait de $f_{X|Y}(\cdot \,|\, y)\, d\mu(\cdot)$ une mesure de probabilité puisque $\{y : f_Y(y) = 0\}$ est $\mathbb{P}_Y$-négligeable. On obtient alors que $f_{X|Y}(\cdot \,|\, \cdot)\, d\mu(\cdot)$ est une version régulière.

A.12 Lois usuelles

A.12.1 Loi gaussienne

Définition A.12.1 (Loi Gaussienne réduite)**.** *Une variable aléatoire X à valeur dans $\mathbb{R}$ est dite gaussienne réduite si sa loi admet pour densité (par rapport à la mesure de Lebesgue sur $\mathbb{R}$) :*

$$g(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$$

La fonction caractéristique associée à loi gaussienne réduite a pour expression :

$$\varphi(t) = \mathbb{E}\left[\exp(itX)\right] = \exp(-t^2/2)$$

Les moments de la loi gaussienne réduite se déduisent du développement de Taylor de $\varphi(t)$ en 0 : les moments d'ordre impair sont nuls et les moments d'ordre pair sont donnés par

$$\mu_{2n} = \mathbb{E}\left[X^{2n}\right] = \frac{(2n)!}{n!\ \ 2^n} = 1 \times 3 \times 5 \ldots \times (2n-1)$$

Définition A.12.2 (Loi gaussienne)**.** *Une variable aléatoire X à valeur dans $\mathbb{R}$ est dite gaussienne si elle peut s'écrire sous la forme $X = \sigma X_r + \mu$ où X_r est une v.a. gaussienne réduite (ce que l'on note sous la forme $X \sim \mathcal{N}(\mu, \sigma^2)$). μ est l'espérance de X et σ^2 sa variance. La densité de X est donnée par :*

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

La fonction caractéristique d'une variable gaussienne de moyenne μ et de variance σ^2 est donnée par

$$\varphi_{\mu,\sigma^2}(t) = \exp\left(i\mu t - \frac{\sigma^2}{2}t^2\right). \tag{A.9}$$

Définition A.12.3 (Loi gaussienne multivariée)**.** *Un vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_n)$ est dit* gaussien *si toute combinaison linéaire $\sum_{j=1}^n \alpha_j X_j = \boldsymbol{\alpha}^T \mathbf{X}$ de ses composantes est une variable aléatoire gaussienne.*

Soit $\mathbf{X} = (X_1, \ldots, X_n)$ un vecteur aléatoire gaussien. Pour $\mathbf{t} = (t_1, \ldots, t_n) \in \mathbb{R}^n$, $Y = \mathbf{t}^T\mathbf{X}$ est une variable gaussienne, dont l'espérance et la variance sont données respectivement par :

$$\mathbb{E}[Y] = \sum_{i=1}^{n} t_i \mathbb{E}[X_i] = \mathbf{t}^T \mathbb{E}[\mathbf{X}]$$

$$\mathbb{E}[(Y - \mathbb{E}[Y])^2] = \sum_{i,j=1}^{n} t_i t_j \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])] = \mathbf{t}^T \mathbf{\Gamma} \mathbf{t}$$

où $\mathbf{\Gamma} = (\mathrm{cov}(X_i, X_j))_{1 \leq i,j \leq n}$ est la matrice de covariance du vecteur $\mathbf{X}$.

Définition A.12.4 (loi $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Gamma})$)**.** *Soit $\boldsymbol{\mu} \in \mathbb{R}^n$ et $\mathbf{\Gamma}$ une matrice semi-définie positive. Nous dirons que $\mathbf{X} = (X_1, \ldots, X_n)$ suit une loi multivariée gaussienne de moyenne $\boldsymbol{\mu}$ et de covariance $\mathbf{\Gamma}$ ($\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Gamma})$), si pour tout $\mathbf{t} = (t_1, \ldots, t_n) \in \mathbb{R}^n$, nous avons $\mathbf{t}^T\mathbf{X} \sim \mathcal{N}(\mathbf{t}^T\boldsymbol{\mu}, \mathbf{t}^T\mathbf{\Gamma}\mathbf{t})$.*

Cette définition implique de façon immédiate :

Proposition A.12.5
Soit $\mathbf{A}$ une matrice $m \times n$ et soit $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Gamma})$. Alors, $\mathbf{AX} + \mathbf{b} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}^T\mathbf{\Gamma}\mathbf{A})$.

Démonstration. Posons $\mathbf{Y} = \mathbf{AX}$ et notons que pour tout $\mathbf{s} \in \mathbb{R}^m$ nous avons :

$$\mathbf{s}^T\mathbf{Y} = (\mathbf{A}^T\mathbf{s})^T\mathbf{X} \sim \mathcal{N}(\mathbf{s}^T\mathbf{A}\boldsymbol{\mu}, \mathbf{s}^T\mathbf{A}\mathbf{\Gamma}\mathbf{A}^T\mathbf{s}).$$

■

Soit $\mathbf{\Gamma} \in \mathbb{R}_n^n$ une matrice symétrique semi-définie positive de rang $k \leq n$. Il existe une matrice $\mathbf{A} \in \mathbb{R}_n^k$ de rang k telle que $\mathbf{\Gamma} = \mathbf{A}\mathbf{A}^T$. Si $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_n)$, nous déduisons de la proposition A.12.5 $\mathbf{AZ} + \boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Gamma})$. Réciproquement, soit $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Gamma})$. Comme $\mathbf{A}$ est de rang k, la matrice $\mathbf{A}^T\mathbf{A} \in \mathbb{R}_k^k$ est inversible et $\mathbf{A}$ est inversible à gauche. Notons $\mathbf{A}^{\#} := (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ son inverse à gauche. Nous avons : $\mathbf{A}^{\#}\mathbf{A} = \mathbf{I}_k$ et $\mathbf{A}\mathbf{A}^{\#}$ est le projecteur orthogonal sur l'image de $\mathbf{A}$ (par construction, $\Im(\mathbf{A}) = \Im(\mathbf{\Gamma})$). Soit $\mathbf{Z} = \mathbf{A}^{\#}(\mathbf{X} - \boldsymbol{\mu})$. La proposition A.12.5 implique que $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_k)$. Nous avons donc :

Proposition A.12.6
Soit $\mathbf{\Gamma} \in \mathbb{R}_n^n$ une matrice semi-définie positive, $\mathrm{rang}(\mathbf{\Gamma}) = k \leq n$ *et soit $\boldsymbol{\mu} \in \mathbb{R}^n$. $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Gamma})$ si et seulement si, pour tout $\mathbf{A} \in \mathbb{R}_n^k$ tel que $\mathbf{A}\mathbf{A}^T = \mathbf{\Gamma}$, il existe $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_n)$ tel que $\mathbf{X} = \mathbf{AZ} + \boldsymbol{\mu}$.*

On pourrait choisir cette caractérisation comme définition de la loi gaussienne $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Gamma})$.
La fonction caractéristique de $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Gamma})$ se déduit directement de (A.9)

$$\varphi_{\boldsymbol{\mu}, \mathbf{\Gamma}}(\mathbf{t}) = \exp(i\mathbf{t}\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^T\mathbf{\Gamma}\mathbf{t}) \tag{A.10}$$

Inversement, si la fonction caractéristique d'une v.a. $\mathbf{X} = (X_1, \ldots, X_n)$ est de la forme A.10, alors pour tout $\mathbf{t} = (t_1, \cdots, t_n)$ et tout $\tau \in \mathbb{R}$:

$$\mathbb{E}[e^{i\tau(\mathbf{t}^T\mathbf{X})}] = \exp\left\{ i\tau\mathbf{t}^T\mu - \frac{\tau^2}{2}\mathbf{t}^T\mathbf{\Gamma}\mathbf{t} \right\}$$

et donc $\mathbf{t}^T\mathbf{X} \sim \mathcal{N}(\mathbf{t}^T\boldsymbol{\mu}, \mathbf{t}^T\mathbf{\Gamma}\mathbf{t})$. Il découle donc de la Définition A.12.3 que $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Gamma})$.
Aussi, puisque la fonction caractéristique caractérise la loi, nous avons :

Proposition A.12.7
$\mathbf{X} = (X_1, \ldots, X_n) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Gamma})$ *si et seulement si sa fonction caractéristique* $\varphi_{\boldsymbol{\mu},\boldsymbol{\Gamma}}(\mathbf{t}) := \mathbb{E}[e^{i\mathbf{t}^T\mathbf{X}}]$ *est donné par :*

$$\varphi_{\boldsymbol{\mu},\boldsymbol{\Gamma}}(\mathbf{t}) = \exp\left(i\mathbf{t}^T\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^T\boldsymbol{\Gamma}\mathbf{t}\right\} \tag{A.11}$$

A.12.2 Propriétés

Soit $n \in \mathbb{N}$ et soit n_1, n_2 tels que $n_1 + n_2 = n$. Pour $\mathbf{x} \in \mathbb{R}^n$, considérons la partition $\mathbf{x} = (\mathbf{x}_1^T, \mathbf{x}_2^T)^T$ avec $\mathbf{x}_1 \in \mathbb{R}^{n_1}$ et $\mathbf{x}_2 \in \mathbb{R}^{n_2}$. De façon similaire, pour $\boldsymbol{\Gamma} \in \mathbb{R}_n^n$ considérons la matrice bloc :

$$\boldsymbol{\Gamma} = \begin{bmatrix} \boldsymbol{\Gamma}_{11} & \boldsymbol{\Gamma}_{12} \\ \boldsymbol{\Gamma}_{21} & \boldsymbol{\Gamma}_{22} \end{bmatrix}.$$

Nous avons :

Proposition A.12.8
Soit $\mathbf{X} = (\mathbf{X}_1^T, \mathbf{X}_2^T)^T \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Gamma})$. $\mathbf{X}_1$ *est indépendant de* $\mathbf{X}_2$ *si et seulement si* $\boldsymbol{\Gamma}_{12} = 0$.

Démonstration. Si $\mathbf{X}_1$ et $\mathbf{X}_2$ sont indépendants, alors $\boldsymbol{\Gamma}_{12} = \mathrm{cov}(\mathbf{X}_1, \mathbf{X}_2) = 0$. Réciproquement, supposons que $\boldsymbol{\Gamma}_{12} = 0$. Comme $\boldsymbol{\Gamma}_{ii}$, $i = 1, 2$ sont semi-définies positives, il existe $\mathbf{A}_i \in \mathbb{R}_{n_i}^{k_i}$ telles que $\boldsymbol{\Gamma}_{ii} = \mathbf{A}_i\mathbf{A}_i^T$, où $k_i = \mathrm{rang}(\boldsymbol{\Gamma}_{ii})$, $i = 1, 2$. Posons :

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & 0 \\ 0 & \mathbf{A}_{22} \end{bmatrix}$$

En utilisant la proposition A.12.6, il existe $\mathbf{Z} \sim \mathcal{N}_{k_1+k_2}(0, \mathbf{I}_{k_1+k_2})$ tel que :

$$\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} = \mathbf{A}\mathbf{Z} + \boldsymbol{\mu} = \begin{pmatrix} \mathbf{A}_{11}\mathbf{Z}_1 + \boldsymbol{\mu}_1 \\ \mathbf{A}_{22}\mathbf{Z}_2 + \boldsymbol{\mu}_2 \end{pmatrix}.$$

Les v.a. $\mathbf{Z}_1$ et $\mathbf{Z}_2$ sont indépendantes car, pour tout $\mathbf{t}_1 \in \mathbb{R}^{n_1}$ et $\mathbf{t}_2 \in \mathbb{R}^{n_2}$ nous avons en vertu de la proposition A.12.7 :

$$\mathbb{E}[e^{i(\mathbf{t}_1^T\mathbf{Z}_1 + \mathbf{t}_2^T\mathbf{Z}_2)}] = \exp(-\|t_1\|^2/2)\exp(-\|t_2\|^2/2) = \mathbb{E}[e^{i(\mathbf{t}_1^T\mathbf{Z}_1)}]\,\mathbb{E}[e^{i(\mathbf{t}_2^T\mathbf{Z}_2)}].$$

Par suite, les v.a. $\mathbf{X}_1$ et $\mathbf{X}_2$ sont indépendantes, ce qui conclut la preuve. ■

Corollaire A.12.9
Soient $\mathbf{A}_1 \in \mathbb{R}_n^{n_1}$ *et* $\mathbf{A}_2 \in \mathbb{R}_n^{n_2}$ *deux matrices telles que* $\mathbf{A}_1^T\mathbf{A}_2 = \mathbf{0}_{n_1 \times n_2}$ *et soit* $\mathbf{Z} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$. *Alors, le vecteur* $\mathbf{Y} = (\mathbf{Y}_1^T, \mathbf{Y}_2^T)^T$ *avec* $\mathbf{Y}_1 := \mathbf{A}_1\mathbf{Z}$ *et* $\mathbf{Y}_2 := \mathbf{A}_2\mathbf{Z}$ *est gaussien et les vecteurs* $\mathbf{Y}_1$ *et* $\mathbf{Y}_2$ *sont indépendants.*

Remarque A.12.10. *La* décorrélation *des composantes d'un vecteur aléatoire n'implique* l'indépendance *de ses composantes que dans le cas où le vecteur est gaussien. Nous donnons un contre-exemple pour illustrer l'importance de cette hypothèse. Soit* X *une v.a. de loi* $\mathcal{N}(0, 1)$ *;* $Y = \epsilon X$, *où* ϵ *est v.a. indépendante de* X *telle que* $\mathbb{P}[\epsilon = 1] = \mathbb{P}[\epsilon = -1] = \frac{1}{2}$. *On démontre aisément que* $Y \sim \mathcal{N}(0, 1)$. *De plus,*

$$\mathbb{E}[XY] = \mathbb{E}[\epsilon X^2] = \mathbb{E}[\epsilon]\mathbb{E}[X^2] = 0,$$

et donc $\mathrm{cov}(X, Y) = 0$. *donc ces v.a. sont décorrélées. Pourtant, elles ne sont pas indépendantes. En effet,* (X, Y) *n'est pas un vecteur aléatoire gaussien puisque* $P[X + Y = 0] = \frac{1}{2}$.

A.12.3 Vecteurs aléatoires gaussiens et densités

Proposition A.12.11
Soit $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Gamma})$ *où* $\boldsymbol{\Gamma}$ *est une matrice définie positive. Alors* $\mathbf{X}$ *possède une densité par rapport à la mesure de Lebesgue sur* $\mathbb{R}^n$ *de la forme :*

$$f_{\boldsymbol{\mu},\boldsymbol{\Gamma}}(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^n} \frac{1}{\sqrt{\det(\boldsymbol{\Gamma})}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Gamma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}, \quad \mathbf{x} \in \mathbb{R}^n. \tag{A.12}$$

Démonstration. Si $\boldsymbol{\Gamma} = \mathbf{I}_n$ la proposition A.12.8 montre que les v.a. $X_1, \ldots, X_n$ sont i.i.d.et donc leur densité jointe est égale au produit des densités marginales, ce qui conduit au résultat dans ce cas particulier. Si $\boldsymbol{\Gamma}$ est une matrice définie positive quelconque, nous utilisons la proposition A.12.6 : il existe $\mathbf{A}$ *inversible* et $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_n)$ tel que $\mathbf{X} = \mathbf{AZ} + \boldsymbol{\mu}$ et l'expression A.12 découle de la formule du changement de variable. ■

La quantité $[(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Gamma}^{-1}(\mathbf{x}-\boldsymbol{\mu})]^{1/2}$ est souvent appelée la *distance de Mahalanobis* de $\mathbf{x}$ à $\boldsymbol{\mu}$. Les lignes de niveaux de la densité $f_{\boldsymbol{\mu},\boldsymbol{\Gamma}}$, *i.e.*les ensembles $\{\mathbf{x} \in \mathbb{R}^n, f_{\boldsymbol{\mu},\boldsymbol{\Gamma}}(\mathbf{x}) = c\}$ correspondent au lieu des points dont la distance de Mahalanobis à $\boldsymbol{\mu}$ est constante. En écrivant $\mathbf{y} = \mathbf{H}^T\mathbf{x}$ où $\mathbf{H}$ est une matrice unitaire qui diagonalise la matrice $\boldsymbol{\Gamma}$, $\mathbf{H}^T\boldsymbol{\Gamma}\mathbf{H} = \mathbf{D}$, $\mathbf{D} = \mathrm{diag}(d_1^2, \ldots, d_n^2)$, les lignes de niveaux sont donc les ellipsoïdes :

$$\sum_{i=1}^{n} (y_i - \nu_i)^2 / d_i$$

centrées en $\boldsymbol{\nu} = \mathbf{H}^T\boldsymbol{\mu}$ et dont les axes principaux sont portés par les vecteurs propres $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_n]$.

A.12.4 Loi Gamma

La loi Gamma est la brique de base permettant de construire de nombreuses autres distributions. La loi Gamma est elle-même liée à la fonction Gamma, définie sur le demi plan complexe $\mathrm{Re}(z) > 0$ par :

$$\Gamma(z) = \int_0^\infty \exp(-t)\ \ t^{z-1} dt = 2\int_0^\infty \exp(-t^2) t^{2z-1} dt. \tag{A.13}$$

En intégrant par partie pour $x > 0$ *réel positif* l'expression précédente, nous avons :

$$\Gamma(x) = [-t^{x-1}e^{-t}]_0^\infty + (x-1)\int_0^\infty t^{x-2}e^{-t}dt = (x-1)\Gamma(x-1)$$

et donc pour n un entier naturel, $\Gamma(n) = (n-1)\Gamma(n-1) = \ldots = (n-1)(n-2)\ldots 1 = (n-1)!$.

Définition A.12.12. *Pour p réel positif, $p > 0$, on appelle loi Gamma réduite à p degrés de liberté (et l'on note $\mathcal{G}amma(p)$) la loi définie sur l'ensemble des réels positifs par la densité*

$$f_p(x) = \Gamma(p)^{-1} \exp(-x)\ \ x^{p-1}, \quad x > 0.$$

Pour $\theta > 0$, on appelle loi Gamma $\mathcal{G}amma(p, \theta)$, la loi de la v.a. $X = \theta Z$, où Z est une loi Gamma à p degrés de liberté (θ est le paramètre d'échelle de la loi).

Si Z est une loi $\mathcal{G}amma(p)$, la définition (A.13) implique que, pour tout $r > -p$, nous avons

$$\mathbb{E}[Z^r] = \frac{\Gamma(p+r)}{\Gamma(p)}, \quad \forall r > -p.$$

Lemme A.12.13
Soit X une v.a. gaussienne centrée réduite. X^2 suit une loi $\mathcal{G}amma(\frac{1}{2}, \frac{1}{2})$.

Démonstration. $\mathbb{P}[X^2 < z] = 0$ si $z < 0$. Pour $z > 0$, nous avons :

$$\begin{aligned}
\mathbb{P}[X^2 < z] &= \mathbb{P}[-\sqrt{z} < X < \sqrt{z}] \\
&= \int_{-\sqrt{z}}^{\sqrt{z}} \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) dx \\
&= 2\int_{0}^{\sqrt{z}} \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) dx \\
&= \frac{2}{\sqrt{2\pi}} \int_0^z \exp(-\frac{u}{2}) \frac{1}{2\sqrt{u}} du
\end{aligned}$$

Ceci conduit au résultat, en utilisant le résultat élémentaire $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. ■

La fonction caractéristique de la loi $\mathcal{G}amma(\theta, p)$ est donnée par :

$$\phi_{\theta,p}(t) = \int_0^\infty \frac{1}{\theta^p \Gamma(p)} x^{p-1} e^{(i\theta t - 1)x/\theta} dx = (1 - i\theta t)^{-p}. \tag{A.14}$$

Cette expression particulière de la fonction caractéristique a pour conséquence immédiate le théorème de convolution suivant pour les lois Gammas.

Lemme A.12.14
Soit $(X_1, \ldots, X_n)$ n v.a. indépendantes distribuées suivant des lois $\mathcal{G}amma(p_i, \theta)$ avec $\theta > 0$ et $p_i > 0$, $i \in \{1, \ldots, n\}$. Alors, $\sum_{i=1}^n X_i$ sont distribuées suivant une loi $\mathcal{G}amma(\sum_{i=1}^n p_i, \theta)$.

A.12.5 Loi du χ^2 à k degrés de liberté

Définition A.12.15 (Loi du χ^2 centrée). *Soient $(X_1, \cdots, X_k)$, k v.a. gaussiennes centrées réduites indépendantes. La v.a. $U = \sum_{i=1}^k X_i^2$ suit une loi appelée loi du χ^2 centrée à k degrés de liberté, notée χ_k^2.*

Proposition A.12.16
La loi du χ_k^2 à k-degrés de liberté est une loi $\mathcal{G}amma(k/2, 1/2)$.

Démonstration. C'est une conséquence immédiate des lemmes A.12.13 et A.12.14. ■

En particulier, pour U une v.a. suivant une loi χ_k^2, nous avons :

$$\mathbb{E}[U] = k \quad \text{et} \quad \text{var}[U] = 2k. \tag{A.15}$$

Définition A.12.17 (Loi non centrée). *Soient $(X_1, \cdots, X_k)$, k v.a. gaussiennes de moyenne μ_i réduites indépendantes. On note $U = \sum_{i=1}^k X_i^2$. On dit que U suit une loi du χ^2 non-centrée à k degrés de liberté, de paramètre de non-centralité $\gamma = (1/2)\sum_{i=1}^k \mu_i^2$; ce que l'on note : $\chi_k^2(\gamma)$.*

Dans la définition ci-dessus, la loi de U ne dépend que de γ, d'où le fait fait que l'on paramétrise la loi par γ seulement, sans avoir à spécifier individuellement les μ_i. Pour le voir, remarquons que, par définition, $U = \|\mathbf{X}+\boldsymbol{\mu}\|_2^2$, où $\mathbf{X}$ est un vecteur Gaussien de composantes indépendantes et $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$. La loi de $\mathbf{X}$ est alors invariante par transformation orthogonale, car si H est une matrice orthogonale, $HX \sim \mathcal{N}(\mathbf{0}, H\mathbf{I}_k H^\top) = \mathcal{N}(\mathbf{0}, \mathbf{I}_k)$, de sorte que $H\mathbf{X} \sim \mathbf{X}$. Ainsi, on a l'égalité en loi, pour toute transformation orthogonale de matrice H,

$$U \stackrel{\text{loi}}{=} \|H\mathbf{X} + \boldsymbol{\mu}\|_2^2 = \|H(\mathbf{X} + H^\top\boldsymbol{\mu})\|_2^2 = \|\mathbf{X} + H^\top\boldsymbol{\mu}\|_2^2$$

Soit H une matrice orthogonale telle que $H^\top\boldsymbol{\mu} = (\|\boldsymbol{\mu}\|_2, 0, \ldots, 0)$. Une telle matrice existe : prendre par exemple la première colonne de H égale à $\frac{1}{\|\boldsymbol{\mu}\|_2}\boldsymbol{\mu}$, puis compléter pour que les colonnes forment une base orthonormale de $\mathbb{R}^k$. On a donc

$$U \stackrel{\text{loi}}{=} \|X + (\|\boldsymbol{\mu}\|_2, 0, \ldots, 0)\| = \|X + (\sqrt{2\gamma}, 0, \ldots, 0)\|,$$

qui ne dépend que de γ.

Proposition A.12.18
La fonction caractéristique d'une de χ^2 à k degrés de liberté et de paramètre de non-centralité γ est donnée par :

$$\left(\frac{1}{\sqrt{1-2\mathrm{i}t}}\right)^k \exp\left(\frac{2\mathrm{i}t\gamma}{1-2\mathrm{i}t}\right)$$

Démonstration. Par définition, si $Z_1, \ldots, Z_{k-1}, X$ sont des v.a. indépendantes, $Z_i \sim \mathcal{N}(0,1)$ et $X \sim \mathcal{N}(\mu, 1)$ alors la v.a. $U = \sum_{i=1}^{k-1} Z_i^2 + X^2 \sim \chi_k^2(\gamma)$ avec $\gamma = \mu^2/2$. Notons que $\sum_{i=1}^{k-1} Z_i^2$ et X^2 sont indépendantes et que $\sum_{i=1}^{k-1} Z_i^2 \sim \chi_{k-1}^2$. Par conséquent,

$$\mathbb{E}[\mathrm{e}^{\mathrm{i}tU}] = (1-2\mathrm{i}t)^{-(k-1)/2}\mathbb{E}[e^{\mathrm{i}tX^2}].$$

Un calcul direct montre que :

$$\begin{aligned}
\mathbb{E}[\mathrm{e}^{\mathrm{i}tX^2}] &= \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}tx^2}(2\pi)^{-1/2}\mathrm{e}^{-(x-\mu)^2/2}dx \\
&= \exp\left[\frac{\mu^2(\mathrm{i}t)}{1-2\mathrm{i}t}\right]\int_{-\infty}^{\infty}(2\pi)^{-1/2}\exp\left(-\frac{1-2\mathrm{i}t}{2}\left[x-\frac{\mu}{1-2\mathrm{i}t}\right]^2\right)dx \\
&= (1-2\mathrm{i}t)^{-1/2}\exp(2\gamma\mathrm{i}t/(1-2\mathrm{i}t)).
\end{aligned}$$

■

Un calcul élémentaire montre que la moyenne et la variance d'une v.a. U distribuée suivant une loi $\chi_k^2(\gamma)$ sont respectivement données par : $\mathbb{E}[\chi^2] = k + \gamma$ et $\mathrm{var}[\chi^2] = 2k + 4\gamma$.

Le résultat suivant joue un rôle important dans la théorie de l'inférence dans les modèles de régression linéaire multiple.

Proposition A.12.19
Soit $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ et soit Π un projecteur orthogonal de rang $k < n$. $\|\Pi\mathbf{Z}\|^2$ est distribué suivant une loi de χ^2 non-centrée à k degrés de liberté de paramètre de non-centralité $\|\Pi\boldsymbol{\mu}\|^2/2$.

Démonstration. Soit $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_k]$ une base orthonormale de l'image de Π. Nous avons donc $\Pi = \mathbf{H}\mathbf{H}^T$ et $\mathbf{H}^T\mathbf{H} = \mathbf{I}_k$ où $\mathbf{I}_k$ est la matrice identité $(k \times k)$. Par conséquent, $\mathbf{Z}^T\Pi\mathbf{Z} = \|\mathbf{e}\|^2$, où $\mathbf{e} = \mathbf{H}^T\mathbf{Z}$. La proposition A.12.5 implique que $\mathbf{e} \sim \mathcal{N}_k(\mathbf{H}^T\boldsymbol{\mu}, \mathbf{I}_k)$. Par conséquent, $\|\mathbf{e}\|^2 \sim \chi_k^2(\delta)$ avec $\delta = \|\mathbf{H}^T\boldsymbol{\mu}\|^2/2 = \boldsymbol{\mu}^T\Pi\boldsymbol{\mu}/2$, ce qui conclut la preuve. ■

A.12.6 Loi de Student

Définition A.12.20. *Soit X et Y deux variables aléatoires indépendantes telles que :*
- *X suit une loi gaussienne centrée réduite,*
- *Y suit une loi du χ^2 centrée à r degrés de liberté,*

Alors $T = X/\sqrt{Y/r}$ suit une loi de Student à r degrés de liberté, que l'on note $\mathbf{T}_r$.

Remarque A.12.21. *"Student" est un pseudonyme utilisé par W.S. Gosset qui, étant employé aux brasseries Guinness, avait besoin de publier sous un nom d'emprunt.*

Proposition A.12.22
La densité d'une loi de Student à r-degrés de liberté est donnée par :

$$f_r(t) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\Gamma\left(\frac{r}{2}\right)} \frac{1}{(r\pi)^{1/2}} \left(1 + \frac{t^2}{r}\right)^{\frac{r+1}{2}}$$

Démonstration. La distribution conjointe des v.a. X et Y est donnée par

$$f_{XY}(x, y) \propto e^{-x^2/2} y^{(r/2)-1} e^{-y/2}, \quad x \in \mathbb{R}, y > 0.$$

En appliquant la transformation $\phi : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R} \times \mathbb{R}^+$, $(x, y) \mapsto (x(y/r)^{-1/2}, y)$, la loi jointe de T et de Y est donnée par :

$$f_{TY}(t, y) = f_{XY}(t(y/r)^{1/2}, y)(y/r)^{1/2}, \quad x \in \mathbb{R}, y > 0,$$

car le Jacobien de la transformation est égal à $(y/r)^{1/2}$. La distribution de T est obtenue en intégrant la loi jointe f_{TY} par rapport à y,

$$f_T(t) \propto \int_0^\infty e^{-y(1+t^2/r)/2} y^{((r+1)/2)-1} dy.$$

et on obtient la formule désirée en faisant le changement de variable $u = y(1 + t^2/r)/2$. ■

Lorsque $r = 1$, la densité de la loi de Student se réduit à

$$f_T(t) = \frac{1}{\pi(1 + t^2)}, \quad t \in \mathbb{R}$$

qui est la densité d'une loi de Cauchy (et donc, qui n'admet pas de moments d'ordre 1). Lorsque $r \to \infty$, le dénominateur (par la loi des grands nombres) tend en probabilité vers 1 et la loi de T tend en loi vers une loi gaussienne standardisée.

Proposition A.12.23
La fonction caractéristique de la loi de Student à r degrés de liberté est donnée par :

$$\phi(t) = \frac{\alpha\pi}{2^{r-1}(\frac{1}{2}(r-1))!} \exp(-|t|\sqrt{r}) \sum_{j=0}^{\frac{1}{2}(r-1)} (2|t|\sqrt{r})^{\frac{1}{2}(r-1)-j} \frac{(\frac{1}{2}(r-1)+j)!}{j!\ (\frac{1}{2}(r-1)-j)!} \tag{A.16}$$

avec $\alpha = \frac{1}{B(\frac{1}{2}, \frac{r}{2})}$. Les moments d'ordre impair sont nuls, et les pairs, qui existent pour $j < \frac{r}{2}$ sont donnés par :

$$\mu_{2j} = \frac{\Gamma(j + \frac{1}{2})\Gamma(\frac{r}{2} - j)}{\Gamma(\frac{1}{2})\Gamma(\frac{r}{2})}$$

Le résultat suivant, du à Gosset (1907), fait partie des "classiques favoris" des statistiques élémentaires et justifie à lui seul l'intérêt porté à la distribution de Student.

Théorème A.12.24
Soit $\mathbf{X} = (X_1, \ldots, X_n) \sim \mathcal{N}(\mu \mathbf{1}_n, \sigma^2 \mathbf{I}_n)$ *où* $\mathbf{1}_n = [1, \ldots, 1]^T$.

1. *Les v.a.* $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ *et* $S^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ *sont indépendantes.*
2. $\bar{X}$ *suit une loi normale* $\mathcal{N}(\mu, \sigma^2/n)$ *et* $(n-1)S^2$ *suit une loi du* χ^2 *à* $(n-1)$ *degrés de liberté.*
3. *La variable* T_n *définie par :*
$$T_n = \sqrt{n} \frac{\bar{X} - \mu}{S}$$
suit une loi de Student à $(n-1)$ *degrés de liberté.*

Démonstration. Notons que $\bar{X} = n^{-1} \mathbf{1}_n^T \mathbf{X}$ et donc que :

$$(n-1)S^2 = \|\mathbf{X} - n^{-1} \mathbf{1}_n \mathbf{1}_n^T \mathbf{X}\|^2 = \|(\mathbf{I}_n - n^{-1} \mathbf{1}_n \mathbf{1}_n^T)\mathbf{X}\|^2.$$

Remarquons que $\Pi := \mathbf{I}_n - n^{-1} \mathbf{1}_n \mathbf{1}_n^T$ est un projecteur orthogonal de rang $(n-1)$ et $\Pi \mathbf{1}_n = 0$. La proposition A.12.19 montre que $(n-1)S^2/\sigma^2$ est distribuée suivant une loi du χ^2 centré à $(n-1)$ degrés de liberté. Le corollaire A.12.9 montre que $\bar{X} = n^{-1} \mathbf{1}_n^T \mathbf{X}$ et $\Pi \mathbf{X}$ sont indépendants et le résultat découle de : $\sqrt{n}(\bar{X} - \mu)/\sigma \sim \mathcal{N}(0, 1)$. ■

Remarque A.12.25. *On peut montrer que la propriété d'indépendance de* $\bar{X}$ *et* S^2 *est* caractéristique *du cas Gaussien : si cette propriété est vérifiée, alors,* $\mathbf{X}$ *est Gaussien.*

A.12.7 Loi de Fisher

Définition A.12.26. *Soient* X *et* Y *deux variables aléatoires indépendantes telles que :*
— X *suit une loi du* χ^2 *centrée à* q*-degrés de liberté,*
— Y *suit une loi du* χ^2 *centrée à* r *degrés de liberté,*
Alors $W = (X/q)/(Y/r)$ *suit une loi de Fisher à* (q, r)*-degrés de liberté, ce que l'on note* $\mathbf{F}(q, r)$.

Proposition A.12.27
La loi de Fisher à (q, r)*-degrés de liberté a une densité donnée par*

$$f(w) = \frac{\Gamma\left(\frac{q+r}{2}\right)}{\Gamma\left(\frac{q}{2}\right)\Gamma\left(\frac{r}{2}\right)} \left(\frac{q}{r}\right)^{q/2} \frac{w^{q/2-1}}{(1 + (q/r)w)^{(q+r)/2}}, \quad w > 0.$$

La preuve est similaire à la preuve de la proposition A.12.22 et est omise. Remarquons que, par définition, si W est distribuée suivant la loi de Fisher $\mathbf{F}_{q,r}$ alors $1/W$ est distribuée suivant la loi de Fisher $\mathbf{F}_{r,q}$. Notons aussi que si T est distribué suivant une loi de Student à r degrés de liberté, alors X^2 est distribué suivant une loi de Fisher à $(1, r)$-degrés de liberté.

Index

Bibliographie

Peter J Bickel and Kjell A Doksum. *Mathematical Statistics : Basic Ideas and Selected Topics, volume I*, volume 117. CRC Press, 2015.

Dominique Foata and Aimé Fuchs. *Calcul des probabilités : cours et exercices corrigés*. Masson, 1996.

E.L. Lehmann. *Testing statistical hypotheses*. John Wiley & Sons, 1959.

Erich Leo Lehmann and George Casella. *Theory of point estimation*, volume 31. Springer Science & Business Media, 1998.

J. Shao. *Mathematical Statistics*. Springer Texts in Statistics. Springer New York, 2008. ISBN 9780387217185.