Notions de probabilités et statistiques

Eric Moisan

eric.moisan@gipsa-lab.grenoble-inp.fr
GIPSA-Lab, ENSE3, 961 Rue de la houille blanche
BP 46, 38402 St Martin d'Hères Cedex

Table des matières

Chapitre 1

Présentation

Dans le monde industriel, l'essentiel des besoins rencontrés par les ingénieurs se réduit à des méthodes de statistiques relativement classiques. Il s'agit généralement d'exploiter ou d'interpréter un nombre parfois important de données. On peut subvenir à ces besoins en se contentant d'appliquer aveuglément les quelques techniques correspondantes. Par contre, dès que l'on sort des sentiers battus, il est hors de question de faire quoi que ce soit sans un minimum de connaissances en probabilités, théorie indispensable à la modélisation de tout problème d'ordre statistique. Par ailleurs, la mécanique quantique et la physique statistique font appel à quelques notions fondamentales des probabilités. Il paraît donc judicieux de prendre le temps d'établir les règles de base de la théorie des probabilités, avant d'aborder, rapidement, les problèmes statistiques les plus fréquents.
Ce cours succinct présente d'abord les notions de base de la théorie des probabilités. Nous effectuons ensuite un rapide tour d'horizon de l'estimation de paramètres. Enfin, nous donnons le principe de base des tests d'hypothèses, sans prendre le temps d'examiner les problèmes les plus fréquents.

Chapitre 2

Probabilités

2.1 Introduction

Tout expérimentateur s'est déjà trouvé confronté au problème suivant : à un moment donné, lors d'une manipulation, on dispose de plusieurs mesures voisines, mais différentes, d'une même quantité. Laquelle faut-il choisir ? On peut être tenté de reproduire l'expérience mais, généralement, la diversité des mesures réapparaît aussitôt et, pire que cela, chaque appareil donne souvent un résultat différent du précédent. Que croire ?
Le remède le plus brutal consiste à considérer la moyenne arithmétique de toutes ces mesures. Une analyse plus fine fera négliger une valeur nettement différente des autres. Enfin, un bon technicien prendra en compte la précision associée à chacune des mesures.
Faute de pouvoir expliquer exactement chacune de ces mesures, il est utile de les modéliser, pour tenir compte de leur aspect imprédictible. C'est là que surgit la théorie des **probabilités**.
Bien assimilées, ces notions permettent alors une amélioration des méthodes d'estimation de la grandeur recherchée. Ce sont les problèmes de **statistiques**.
Enfin, des techniques plus récentes d'estimation tentent de prendre en compte le fait que certaines hypothèses introduites lors de la modélisation probabiliste d'un problème ne sont pas toujours vérifiées. Comment faire pour que les résultats statistiques conservent un certain crédit ? Ce sont les problèmes de **robustesse**.

2.1.1 Origines de la théorie des probabilités

Au $17^{ème}$ siècle, Pascal (1623-1662) et Fermat (1601-1665) se sont intéressés aux jeux de hasard. Ils ont ainsi introduit le dénombrement. Mais il faut attendre Kolmogorov (1903-1987) pour formaliser complètement la théorie des probabilités, telle qu'elle apparaît maintenant. Cette modélisation moderne s'appuie sur l'algèbre de Boole.

2.1.2 Notions intuitives

Pour des expériences simples, chacun intuite volontiers la notion de probabilité (exemple : probabilité d'obtenir un chiffre donné lors du lancer d'un dé équilibré). En fait, derrière cette notion, se cache une intuition expérimentale. On lance N fois un dé équilibré à six faces et on compte le nombre de cas où l'on obtient le chiffre 4, par exemple. Soit n(N) ce nombre entier. On définit la *fréquence relative d'apparition*, $n(N)/N$, que l'on peut tracer en fonction de N. Si on fait tendre le nombre d'expériences vers l'infini, cette quantité tend vers une limite, égale à 1/6, qui n'est autre que la probabilité d'obtenir le résultat voulu sur une seule épreuve.

2.2 Mesures et probabilités

Intuitivement, la "probabilisation" d'un ensemble fini d'événements est simple (dé équilibré). Elle est également suffisante pour satisfaire les règles que nous allons bientôt décrire. Mais comment faire quand il y a une infinité (dénombrable ou non) d'événements ? (temps de vol d'un électron). Il faut se référer à la théorie de la mesure...

2.2.1 Définitions

La mise en place des probabilités sur l'ensemble Ω des *événements élémentaires* ω d'une *épreuve* se fait en deux étapes.
Tout d'abord, il faut construire une *tribu*, notée $\mathcal{A}$, obtenue en groupant certaines parties de Ω, notées A_i et appelées *événements*, selon les règles suivantes :
– $\Omega \in \mathcal{A}$
– $\forall A_i \in \mathcal{A}, \quad \complement_\Omega\ A_i\ \in \mathcal{A}$
– $\cup A_i \in \mathcal{A}$
En d'autres termes, $\mathcal{A}$ est stable par union et intersection. Le doublet $(\Omega, \mathcal{A})$ est un espace qualifié de *probabilisable*.
Pour concrétiser ces notions, examinons les au travers d'un exemple simple. Considérons *l'épreuve* constituée du lancer d'un dé équilibré. Les *événements élémentaires* étant les six chiffres de 1 à 6, l'ensemble Ω s'écrit $\{1,2,3,4,5,6\}$. Si on ne s'intéresse qu'à la parité du résultat, on peut se contenter de construire la *tribu* engendrée par le sous-ensemble $\{2,4,6\}$. Cette tribu s'écrit de façon exhaustive $\mathcal{A} = \left\{\ \Omega, \emptyset, \{2,4,6\}, \{1,3,5\}\ \right\}$.
La seconde étape consiste en l'attribution d'une mesure de probabilité à chaque élément A_i de la tribu $\mathcal{A}$. Cette mesure, notée p, n'est rien d'autre qu'une application de $\mathcal{A}$ dans [0,1], qui doit vérifier en outre :
– $p(\Omega) = 1$
– $p(\cup A_i) = \sum p(A_i)$ pour peu que les A_i soient disjoints deux à deux (σ *additivité*)
Le triplet $(\Omega, \mathcal{A}, p)$ est un espace qualifié de *probabilisé*.

2.2.2 Propriétés de la mesure de probabilité

A_i, A_k désignant des éléments de la tribu $\mathcal{A}$, construite sur Ω :
– $p(\emptyset) = 0 \leq p(A_i) \leq p(\Omega) = 1$
– $p(\complement_\Omega\ A) = 1 - p(A)$
– $p(A_i \cup A_k) = p(A_i) + p(A_k) - p(A_i \cap A_k)$
Il faut soustraire $p(A_i \cap A_k)$ car cette quantité est éventuellement mesurée deux fois, au travers de $p(A_i)$ d'une part et $p(A_k)$ d'autre part.

2.2.3 Indépendance - Probabilité conditionnelle

Deux événements A et B sont *indépendants* si et seulement si $p(A \cap B) = p(A)\ p(B)$.
On note $p(A \mid B)$ la *probabilité conditionnelle* de A sachant B (probabilité de réaliser l'événement A, sachant que l'événement B est lui-même réalisé).
Intuitivement, pour un ensemble de N expériences réalisées, on a $p(A) = n_A/N$, $p(B) = n_B/N$ et $p(A,B) = n_{AB}/N$. On peut alors approcher la probabilité de A sachant B par n_{AB}/n_B. D'où la définition de la probabilité conditionnelle

$$p(A \mid B) \triangleq \frac{p(A \cap B)}{p(B)} \tag{2.1}$$

Cette quantité est bien inférieure ou égale à 1 pour peu que l'on explicite le dénominateur

$$p(B) = p\Big((B \cap A) \ \cup \ (B \cap \complement_\Omega \ A)\Big) \ = \ p(B \cap A) + p(B \cap \complement_\Omega \ A) \ - \ p(\emptyset)$$

Notons les trois cas particuliers :
– A et B indépendants $\Rightarrow$ $p(A \mid B) = p(A)$
– A et B incompatibles $\Rightarrow$ $p(A \mid B) = 0$ puisque $A \cap B = \emptyset$
– $B \subset A$ $\Rightarrow$ $p(A \mid B) = 1$ puisque $A \cap B = B$

2.2.4 Probabilités totales

Soit une partition $\{A_i\}_{i\in\mathbb{N}}$ de $\Omega : \cup A_i = \Omega$ et $A_i \cap A_k = \emptyset, \quad \forall k \neq i$.
On dit que les A_i forment un système complet d'événements. Considérons également un événement quelconque B. L'ensemble des intersections $A_i \cap B$ constitue alors une partition de B lui-même. On peut ainsi exprimer la probabilité de cet événement B selon

$$\forall B \in \mathcal{A} \qquad p(B) = \sum_i p(B \mid A_i) \ p(A_i) \tag{2.2}$$

2.2.5 Bayes

Cette formule exploite la précédente pour exprimer la probabilité d'un premier événement conditionné par un second, en fonction de la probabilité conditionnelle du second sachant le premier. Soit une partition $\{A_i\}_{i\in\mathbb{N}}$ de $\Omega : \cup A_i = \Omega$ et $A_i \cap A_k = \emptyset, \quad \forall k \neq i$. On a

$$p(A_k \mid B) = \frac{p(B \mid A_k)p(A_k)}{\sum_i p(B \mid A_i) \ p(A_i)} \tag{2.3}$$

En effet, le numérateur n'est autre que la mesure de probabilité de l'intersection de A_k et B, tandis que le dénominateur, en vertu de la loi des probabilités totales, est égal à $p(B)$.

2.3 Variable aléatoire

Par définition, une application X, de $(\Omega, \mathcal{A})$ dans $(\Omega', \mathcal{A}')$, est qualifiée de **mesurable** si :

$$\forall A' \in \mathcal{A}' \quad X^{-1}[A'] = A \in \mathcal{A}$$

Une variable aléatoire, à valeurs dans Ω', est une application mesurable dont on a probabilisé l'ensemble de départ Ω, à l'aide d'une mesure de probabilité, notée p.
Dans toute la suite, on s'intéressera essentiellement aux variables aléatoires réelles, pour lesquelles $\Omega' = \mathbb{R}$ et $\mathcal{A}'$ est la tribu borélienne (tribu engendrée par les semi-ouverts $]-\infty, x]$).(Borel 1871-1956).

2.3.1 Variable aléatoire réelle

X est une application mesurable, à valeurs dans $\mathbb{R}$ muni de la tribu Borélienne. On peut donc, sans ambiguité, associer une mesure de probabilité à chaque élément de cette tribu, entre autres, à chaque intervalle du type $]-\infty, x]$.
On appelle fonction de répartition, notée F_X, la mesure de probabilité associée à ces intervalles. Cette fonction est définie par :

$$\begin{aligned} F_X \ : \quad \mathbb{R} \ &\to [0,1] \\ x \ &\to F_X(x) \triangleq \text{proba}\Big\{X \in]-\infty, x]\Big\} = p\ \Big\{X^{-1}(\ ]-\infty, x]\ )\Big\} \end{aligned} \tag{2.4}$$

Elle possède les propriétés suivantes :
- $\lim_{x\to-\infty} F_X(x) = 0$
- $\lim_{x\to+\infty} F_X(x) = 1$
- $\text{proba}\{a < X \leq b\} = \begin{cases} F_X(b) - F_X(a) & \text{si } b > a, \\ 0 & \text{sinon.} \end{cases}$
- elle est croissante et continue à droite en tout point

2.3.1.1 Variable aléatoire réelle continue

C'est une variable aléatoire à valeurs dans $\mathbb{R}$, dont la fonction de répartition est continue. On définit alors la densité de probabilité, notée f_X, comme étant une fonction positive telle que :

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du \tag{2.5}$$

En voici trois exemples parmi les plus courants :

– loi uniforme	$f_X(x) = \frac{u(x-a)-u(x-b)}{b-a}$	$F_X(x) = \frac{x-a}{b-a}$	$a \leq x \leq b$
– loi de Gauss	$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-m)^2}{2\sigma^2})$	$F_X(x) = \frac{1}{2}\left\{1 + \text{Erf}(\frac{x-m}{\sigma\sqrt{2}})\right\}$	$\sigma > 0$ [1]
– loi de Cauchy	$f_X(x) = \frac{a}{\pi(a^2+x^2)}$	$F_X(x) = \frac{1}{2} + \frac{1}{\pi} \arctan(\frac{x}{a})$	$a > 0$

2.3.1.2 Variable aléatoire réelle discrète

C'est une variable aléatoire dont l'ensemble des valeurs est fini ou dénombrable. En conséquence, sa fonction de répartition est une fonction en escaliers. On définit malgré tout une densité de probabilité, toujours notée f_X, à l'aide des distributions de Dirac :

$$f_X(x) = \sum_i \text{proba}(X = x_i)\ \delta_{x_i} \tag{2.6}$$

Voici les exemples les plus fréquents :

– loi de Bernouilli	$\text{proba}(X = x_1) = p$ et $\text{proba}(X = x_2) = q$	2 états $\Rightarrow$ $p+q = 1$
– loi binômiale	$\text{proba}(X = k) = \text{C}_n^k\ p^k\ (1-p)^{n-k}$	$0 \leq k \leq n$ [2]
– loi de Poisson	$\text{proba}(X = k) = \frac{\lambda^k}{k!}\ e^{-\lambda}$	$\lambda > 0$ et $k \in \mathbb{N}$

On montre aisément que la somme de n variables aléatoires de Bernouilli de même paramètre p, deux à deux indépendantes, suit une loi binômiale, de paramètres (n, p).
D'autre part, nous verrons par la suite que, lorsque n tend vers l'infini, une binômiale (n, p) peut s'approcher par une gaussienne (cf. paragraphe 2.8.3) ou par une loi de Poisson de paramètre $\lambda = np$, pour peu que p soit faible (inférieur à 0,1).

2.3.2 Variable aléatoire vectorielle

$\Omega' = \mathbb{R}^n$ et $\mathcal{A}'$ est la tribu engendrée par les semi-ouverts $]-\infty, x_1]$x ... x $]-\infty, x_n]$. L'extension du cas scalaire ne présente pas de difficultés particulières, hormis des problèmes de notation. Le cas échéant, afin d'alléger cette dernière, on désignera par $\underline{X}$ (resp. $\underline{x}$) le vecteur composé de $X_1, X_2, \ldots, X_n$ (resp. $x_1, \ldots, x_n$). La fonction de répartition est ainsi définie par :

$$\begin{aligned} F_{\underline{X}} : \quad \mathbb{R}^n &\to [0,1] \\ \underline{x} &\to F_{\underline{X}}(\underline{x}) \triangleq \text{proba}\Big\{X_1 \in]-\infty, x_1], ..., X_n \in]-\infty, x_n]\Big\} \end{aligned} \tag{2.7}$$

1. la fonction Erf est définie comme $Erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-u^2} du$
2. $\text{C}_n^k = \frac{n!}{k!\ (n-k)!}$ désigne le nombre de combinaisons de k éléments parmi n.

Elle possède les propriétés suivantes :
- $F_{X_1,...,X_n}(-\infty, x_2, ..., x_n) = \cdots = F_{X_1,...,X_n}(x_1, ..., x_{n-1}, -\infty) = 0$
- $F_{X_1,...,X_n}(+\infty, ..., +\infty) = 1$
- $\text{proba}\{a_1 < X_1 \leq b_1,\ a_2 < X_2 \leq b_2\} =$
$= F_{X_1,X_2}(b_1, b_2) - F_{X_1,X_2}(b_1, a_2) - F_{X_1,X_2}(a_1, b_2) + F_{X_1,X_2}(a_1, a_2)$
- elle est positive et continue à droite en tout point

On définit également une densité de probabilité, $f_{\underline{X}}(\underline{x})$, telle que :

$$\forall (x_1, ..., x_n) \qquad F_{X_1,...,X_n}(x_1, ..., x_n) = \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_n} f_{X_1,...,X_n}(u_1, ..., u_n)\ du_1 ... du_n$$

soit, dans le cas général

$$f_{X_1,...,X_n}(x_1, ..., x_n) = \frac{\partial^n}{\partial x_1 ... \partial x_n}\ F_{X_1,...,X_n}(x_1, ..., x_n) \quad \geq 0 \tag{2.8}$$

Enfin, la connaissance de la loi de la variable vectorielle permet de retrouver celle d'un sous-ensemble de $\underline{X}$, par le biais des **lois marginales** : $F_{X_1}(x_1) = F_{X_1,X_2,...,X_n}(x_1, +\infty, ..., +\infty)$
soit, en termes de densité de probabilité :

$$f_{X_1}(x_1) = \int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} f_{X_1,...,X_n}(x_1, x_2, ..., x_n)\ dx_2 ... dx_n$$

2.3.3 Lois conditionnelles

Conformément à la formule de Bayes (équ.(2.3)), on peut construire la fonction de répartition conditionnelle :

$$\begin{aligned} \text{proba}\Big(Y \in]-\infty, y] \mid X \in]x, a+x]\Big) &= \frac{\text{proba}\{Y \in]-\infty, y], X \in]x, a+x]\}}{\text{proba}\{X \in]x, a+x]\}} \\ F_Y\Big(y \mid x < X \leq a+x\Big) &= \frac{\int_x^{a+x} du \int_{-\infty}^{y} f_{X,Y}(u, v)\ dv}{\int_x^{a+x} du \int_{-\infty}^{+\infty} f_{X,Y}(u, v)\ dv} \end{aligned}$$

A la limite, si on fait tendre a vers 0, le numérateur et le dénominateur tendent tous deux vers 0, mais leur rapport se comporte comme

$$\begin{aligned} F_Y(y \mid X \approx x) &= \frac{\int_{-\infty}^{y} f_{X,Y}(x, v)\ dv}{\int_{-\infty}^{+\infty} f_{X,Y}(x, v)\ dv} \\ &= \frac{\int_{-\infty}^{y} f_{X,Y}(x, v)\ dv}{f_X(x)} \\ &= \int_{-\infty}^{y} \frac{f_{X,Y}(x, v)}{f_X(x)}\ dv \end{aligned}$$

d'où l'on déduit la définition de la **densité de probabilité conditionnelle**, pour peu que la fonction de répartition précédente soit différentiable par rapport à la variable y

$$f_Y(y \mid x) \triangleq \frac{f_{X,Y}(x, y)}{f_X(x)} \tag{2.9}$$

Il en découle la formule dite des **probabilités totales**

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y)\ dx = \int_{-\infty}^{+\infty} f_Y(y \mid x)\ f_X(x)\ dx \tag{2.10}$$

2.3.4 Variables aléatoires indépendantes

Deux variables aléatoires X et Y sont indépendantes si, pour tout couple (x, y), les événements $X \in]-\infty, x]$ et $Y \in]-\infty, y]$ le sont, c'est à dire si $F_{X,Y}(x, y) = F_X(x)\ F_Y(y)$, voire en termes de densité de probabilité $f_{X,Y}(x, y) = f_X(x)\ f_Y(y)$.
En conséquence, la densité de probabilité conditionnelle de deux variables X et Y indépendantes n'est autre que $f_Y(y \mid x) = f_Y(y)$.

2.4 Valeurs moyennes

Lors d'une expérience, pour estimer une quantité inconnue à partir d'un ensemble de n mesures voisines, mais différentes, on calcule souvent la moyenne arithmétique de ces n valeurs. D'un point de vue probabiliste, cela revient à dire que l'on considère chaque mesure m_i (i variant de 1 à n) comme une réalisation possible d'une même variable aléatoire M, dotée d'une probabilité constante, égale à $1/n$. Dans ce cas, la moyenne arithmétique coïncide avec **l'espérance mathématique** de la variable aléatoire M.

2.4.1 Espérance mathématique d'une variable aléatoire

Dans le cas d'une variable aléatoire X discrète, il est naturel de considérer sa moyenne, notée $E(X)$, comme égale à $\sum_i x_i\ \text{proba}(X = x_i)$. Intuitivement, on peut étendre cette définition au cas d'une variable aléatoire continue :
$E(X) = \lim_{h\to 0} \sum_i ih\ \text{proba}\{(i-1)h < X \leq ih\} = \lim_{h\to 0} \sum_i ih\ \{F_X(ih) - F_X((i-1)h)\}$.
Il convient de noter que cette quantité peut ne pas exister, pour certaines fonctions de répartition.

2.4.1.1 Définition

X étant une variable aléatoire de $(\Omega, \mathcal{A}, p)$ dans $(\Omega', \mathcal{A}')$, supposée p-intégrable, on définit son espérance mathématique comme

$$E(X) \triangleq \int_\Omega X(\omega)\ dp(\omega)$$

2.4.1.2 Propriétés

L'opérateur espérance mathématique est linéaire, donc $E(aX + b) = aE(X) + b$.
De plus, quelle que soit g, fonction mesurable, $g(X)$ est une variable aléatoire dont la moyenne, si elle existe, est égale à $E(g[X]) = \int_\Omega g[X(\omega)]\ dp(\omega)$.

2.4.1.3 Lien avec la fonction de répartition

Cette définition n'étant pas très parlante, on lui préfere souvent sa "traduction" à l'aide de la fonction de répartition :

$$E(X) = \int_{\Omega'} x\ dF_X(x) \tag{2.11}$$

Ou mieux encore, à l'aide de la densité de probabilité, dans le cas d'une v. a. continue :

$$E(X) = \int_{-\infty}^{+\infty} x\ f_X(x)\ dx \tag{2.12}$$

ou bien dans le cas d'une variable aléatoire discrète

$$E(X) = < f_X, Id > \tag{2.13}$$

2.4.2 Moments d'une variable aléatoire réelle

2.4.2.1 Moments d'ordre n

Ce sont les espérances mathématiques des variables X^n : $E(X^n)$, $n \in \mathbb{N}^*$. Ces quantités n'existent pas pour tout n. Mais si le moment d'ordre n_0 existe, alors tous les moments d'ordre inférieur à n_0 existent également.

2.4.2.2 Moments centrés

Ce sont les moments de la variable aléatoire centrée (de moyenne nulle) $X_c = X - E(X)$.

2.4.2.3 Variance et écart-type

Ces quantités mesurent la dispersion d'une variable aléatoire autour de sa moyenne. La variance désigne l'écart quadratique moyen entre X et $E(X)$:

$$Var(X) \triangleq E(X_c^2) \geq 0. \tag{2.14}$$

On vérifie aisément $Var(aX + b) = a^2\ Var(X)$.
L'écart-type n'est autre que la racine carrée de la variance, afin de manipuler une quantité homogène à la variable :

$$\sigma \triangleq \sqrt{Var(X)} \tag{2.15}$$

On définit la variable aléatoire centrée réduite, $Y = \frac{X - E(X)}{\sigma}$, qui est de moyenne nulle (centrée) et de variance unitaire (réduite).
En pratique, on calcule souvent la variance par application du théorème de Koenig, dont la démonstration est évidente : $Var(X) = E(X^2) - E^2(X)$.

2.4.2.4 Inégalité de Bienaymé-Tchebichev

$$\text{proba}(\ |\ X - E(X)\ |\ > a)\ \leq\ \frac{\sigma^2}{a^2} \tag{2.16}$$

où a désigne un réel positif.
Cette inégalité, qui résulte de deux majorations successives, est relativement grossière.
Soit Δ l'ensemble des valeurs de x pour lesquelles $|\ X - E(X)\ | > a$.

$$\begin{aligned}
\text{proba}(|\ X - E(X)\ | > a) &= \int_{\Delta} f_X(x)\ dx \\
&\leq \int_{\Delta} \Big(\frac{X - E(X)}{a}\Big)^2\ f_X(x)\ dx \\
&\leq \int_{\mathbb{R}} \Big(\frac{X - E(X)}{a}\Big)^2\ f_X(x)\ dx \\
&= \frac{1}{a^2}\ Var(X)
\end{aligned}$$

La première inégalité résulte du fait que l'on multiplie l'intégrande positive par un terme supérieur ou égal à 1 ($\frac{X - E(X)}{a}$). La seconde découle de l'addition d'une quantité positive, liée à la probabilité qu'a la variable aléatoire X de ne pas prendre ses valeurs dans Δ.

2.4.3 Cas particulier de la gaussienne

Parmi l'ensemble des variables aléatoires, la plus fréquente est sans conteste la gaussienne, dont la loi ne dépend que de deux paramètres, sa moyenne ($m \in \mathbb{R}$) et son écart-type ($\sigma \geq 0$).

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-m)^2}{2\sigma^2})$$

La moyenne se calcule selon

$$\begin{aligned}
E(X) &= \int_{-\infty}^{+\infty} x\ f_X(x)\ dx \\
&= \int_{-\infty}^{+\infty} x\ \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-m)^2}{2\sigma^2})\ dx \\
&= \int_{-\infty}^{+\infty} (m+\sigma y)\ \frac{1}{\sqrt{2\pi}} \exp(-\frac{y^2}{2})\ dy \qquad (y = \frac{x-m}{\sigma},\ \sigma > 0) \\
&= m\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{y^2}{2})\ dy\ +\ \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} y\ \exp(-\frac{y^2}{2})\ dy \\
&= m(F_Y(+\infty) - F_Y(-\infty))\ +\ \frac{\sigma}{\sqrt{2\pi}} \Big[\exp(-\frac{y^2}{2})\Big]_{+\infty}^{-\infty} \\
&= m\ (1-0)\ +\ \frac{\sigma}{\sqrt{2\pi}}\ 0 \\
E(X) &= m
\end{aligned} \tag{2.17}$$

Quant à la variance

$$\begin{aligned}
E\{(X-m)^2\} &= \int_{-\infty}^{+\infty} (x-m)^2\ \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-m)^2}{2\sigma^2})\ dx \\
&= \int_{-\infty}^{+\infty} \sigma^2 y^2\ \frac{1}{\sqrt{2\pi}} \exp(-\frac{y^2}{2})\ dy \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \big[-y\ \exp(-\frac{y^2}{2})\big]_{-\infty}^{+\infty}\ +\ \sigma^2 \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{y^2}{2})\ dy \\
&= 0\ +\ \sigma^2(F_Y(+\infty) - F_Y(-\infty)) \\
\mathrm{Var}(X) &= \sigma^2
\end{aligned} \tag{2.18}$$

On écrit généralement que X suit $\mathcal{N}(m, \sigma^2)$.
Enfin, la variable centrée réduite, $Y = \frac{X-m}{\sigma}$, suit une loi normale $\mathcal{N}(0,1)$.

2.4.4 Moments d'un couple de variables aléatoires

Soient $\underline{X} = (X_1, \ldots, X_n)$ une v. a. vectorielle et h une fonction mesurable de $\mathbb{R}^n$ dans $\mathbb{R}$. Alors $h(\underline{X})$ est une v. a. réelle, dont la moyenne, si elle existe, est donnée par

$$E\{h(\underline{X})\} = \int \ldots \int_{\mathbb{R}^n} h(\underline{x})\ f_{\underline{X}}(\underline{x})\ d\underline{x} \tag{2.19}$$

En particulier, on définit les moments d'un couple (X,Y) de v. a. par $E(X^j\ Y^k)$, dont le **moment croisé du second ordre**, $E(XY)$, qui joue un rôle privilégié quant à l'analyse des liens entre X et Y.
Théorème :
Si $E(X^2)$ et $E(Y^2)$ existent, alors $E(XY)$ existe et vérifie l'inégalité $E^2(XY) \leq\ E(X^2)\ E(Y^2)$.

La démonstration repose sur l'étude du moment du second ordre de la v. a. $Z = X + \alpha Y$, où α est un paramètre réel

$$E[(X + \alpha Y)^2] = \alpha^2 E(Y^2) + 2\alpha E(XY) + E(X^2)$$

Cette quantité étant positive, quel que soit α, le discriminant de cette expression du second degré en α est forcément négatif.

$$E^2(XY) - E(Y^2)\ E(X^2)\ \leq\ 0$$

En outre, ce moment croisé possède les propriétés suivantes
- $E(XY) = E\Big\{[X_c + E(X)]\ [Y_c + E(Y)]\Big\} = E(X_c Y_c) + E(X)E(Y)$
- X et Y indépendantes $\Rightarrow\ E(XY) = E(X)E(Y)$ et donc $E(X_c Y_c) = 0$.
  Deux variables indépendantes sont décorrélées. Attention, la réciproque est fausse ! décorrélation $\not\Rightarrow$ indépendance (à l'exception du cas particulier gaussien).
- X et Y indépendantes $\Rightarrow Var(X + Y) = Var(X) + Var(Y)$.
  En effet, $Var(X + Y) = E\Big\{(X_c + Y_c)^2\Big\} = E(X_c^2 + Y_c^2 + 2X_c Y_c) = E(X_c^2) + E(Y_c^2) + 0$

On appelle **covariance** le moment croisé du second ordre des variables centrées :

$$\mathrm{Cov}(X, Y) = E(X_c Y_c) \tag{2.20}$$

et **coefficient de corrélation**, noté r, cette quantité normée par le produit des écarts-types

$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} \tag{2.21}$$

On vérifie aisément que
- $|\ \mathrm{Cov}(X, Y)\ | \leq\ \sigma_X\ \sigma_Y$ ce qui entraine $|\ r\ | \leq\ 1$
- $|\ r\ | = 1$ signifie que X et Y sont linéairement dépendantes.
- X et Y indépendantes $\Rightarrow r = 0$. (la réciproque est fausse, hors mis le cas particulier gaussien).

L'ensemble des v. a. qui possèdent une variance constitue un espace vectoriel sur $\mathbb{R}$, que l'on peut munir d'un produit scalaire : $E(XY)$. C'est bien une forme bilinéaire symétrique positive : $E(X^2) = 0 \ \Rightarrow\ X = 0$ en moyenne quadratique, puisque $E(X^2) = 0 \ \Rightarrow\ E(X) = 0$ et $\mathrm{Var}(X) = 0$.

2.4.5 Espérance conditionnelle

Techniquement, pour calculer une espérance mathématique, il peut s'avérer utile de définir une notion propre aux variables vectorielles.
Pour alléger les notations, considérons une variable aléatoire de dimension 2, notée (X, Y).
Conformément à l'équation (2.19), la moyenne de la seule composante X d'une telle variable s'écrit

$$\begin{aligned}
E(X) &= \int\int_{\mathbb{R}^2} x\ f_{X,Y}(x, y)\ dxdy \\
&= \int\int_{\mathbb{R}^2} x\ f_{X|Y}(x|y)\ f_Y(y)\ dxdy \\
&= \int_{\mathbb{R}} \underbrace{\Big[\int_{\mathbb{R}} x\ f_{X|Y}(x|y)\ dx\Big]}_{E(X|Y)}\ f_Y(y)dy \\
&= \int_{\mathbb{R}} E(X|Y)\ f_Y(y)dy
\end{aligned}$$

Le terme $E(X|Y)$ représente l'espérance mathématique de la variable X, conditionnée par le fait que la variable Y a pris y pour réalisation. C'est une nouvelle variable aléatoire, que l'on peut interpréter comme une fonction de Y et dont la densité de probabilité découle donc de f_Y.
La notion correspondante prend tout son sens lorsque l'on calcule les courbes de régression (cf. paragraphe 2.7).
Cette définition se généralise sans peine au cas où les variables X et Y sont elles-mêmes vectorielles.

2.5 Changement de variables

Considérons $\underline{X}$ une variable aléatoire définie sur $(\Omega, \mathcal{A}, p)$ dans $\mathbb{R}^n$. Soit h une application mesurable de $\mathbb{R}^n$ dans $\mathbb{R}^n$. On sait maintenant que $\underline{Y} = h(\underline{X})$ est une variable aléatoire vectorielle. Comment déduire la densité de probabilité $f_{\underline{Y}}$ de $\underline{Y}$, de $f_{\underline{X}}$, densité de $\underline{X}$? Nous allons répondre progressivement à cette question, en partant du cas le plus simple.

2.5.1 Dimension 1

Supposons la fonction h bijective, croissante et dérivable. La probabilité pour que Y prenne ses valeurs dans un intervalle $]-\infty, y]$ est donnée par la probabilité pour que X appartienne à $]-\infty, h^{-1}(y)]$. Ceci s'écrit directement à l'aide des fonctions de répartition.

$$F_Y(y) = F_X[h^{-1}(y)]$$

La densité de probabilité s'obtient par simple dérivation selon y.

$$\begin{aligned} f_Y(y) &= f_X[h^{-1}(y)]\ \frac{dx}{dy} \\ &= f_X[h^{-1}(y)]\ \frac{1}{h'[h^{-1}(y)]}\ \geq\ 0 \end{aligned}$$

Le calcul varie très peu lorsque la fonction h , toujours bijective et dérivable, est décroissante. La seule différence est que l'antécédent de l'événement $Y \in]-\infty, y]$ est maintenant $X \in [h^{-1}(y), +\infty[$, ce qui se traduit par :

$$\begin{aligned} F_Y(y) &= 1 - F_X[h^{-1}(y)] \\ f_Y(y) &= -f_X[h^{-1}(y)]\ \frac{dx}{dy} \\ &= f_X[h^{-1}(y)]\ \frac{-1}{h'[h^{-1}(y)]}\ \geq\ 0 \end{aligned}$$

Ces deux cas particuliers peuvent s'écrire selon une formule unique :

$$f_Y(y) = f_X\Big(h^{-1}(y)\Big)\ \left|\frac{1}{h'[h^{-1}(y)]}\right| \tag{2.22}$$

Dans le cas général où la fonction h est dérivable mais pas bijective, il suffit de partitionner $\mathbb{R}$ en sous-ensembles D_i sur lesquels h est monotone. Ces intervalles étant disjoints, on obtient une somme des mesures de probabilité correspondantes.

$$f_Y(y) = \sum_{i \in I}\ f_X(x_i) \left|\frac{dx_i}{dy_i}\right| \quad , \quad I = \{i \in \mathbb{N}\ /\ h^{-1}(y) \in D_i\} \tag{2.23}$$

Exemple : $Y = a \sin X$, où X est une v.a. uniforme sur $]\frac{-\pi}{2}, \frac{3\pi}{2}]$, tandis que a est une amplitude donnée. Cette fonction n'étant pas bijective, il faut considérer séparément les deux intervalles $D_1 =]\frac{-\pi}{2}, \frac{\pi}{2}]$ ($y(x)$ y est une fonction croissante) et $D_2 =]\frac{\pi}{2}, \frac{3\pi}{2}]$ (fonction décroissante). A y donné dans $[-a, a]$ correspondent deux antécédents $x_1 = \arcsin \frac{y}{a}$ et $x_2 = \pi - x_1$.

$$\begin{aligned}
F_Y(y) &= \text{proba}(X \leq x_1) \ + \ \text{proba}(X \geq x_2) \\
&= F_X(x_1) \ + \ 1 - F_X(x_2) \\
f_Y(y) &= f_X\Big(\arcsin(\frac{y}{a})\Big) \ \frac{d}{dy}\Big(\arcsin(\frac{y}{a})\Big) \ - \ f_X\Big(\pi - \arcsin(\frac{y}{a})\Big) \ \frac{d}{dy}\Big(\pi - \arcsin(\frac{y}{a})\Big) \\
&= \Big\{ f_X\Big(\arcsin(\frac{y}{a})\Big) \ + \ f_X\Big(\pi - \arcsin(\frac{y}{a})\Big) \Big\} \ \frac{d}{dy}\Big(\arcsin(\frac{y}{a})\Big) \\
&= \frac{1+1}{2\pi} \ \frac{1}{\sqrt{a^2 - y^2}} \\
&= \frac{1}{\pi\sqrt{a^2 - y^2}} \qquad \text{pour } y \in \ [-a, a]
\end{aligned}$$

2.5.2 Dimension 2

On considère maintenant un couple (X, Y), de densité de probabilité $f_{X,Y}$, auquel la fonction h associe le couple (U, V) : $\qquad (U, V) = h(X, Y)$.
Théorème : si h est bijective et continuement différentiable, le couple (U, V) admet une d.d.p.

$$f_{U,V}(u, v) = f_{X,Y}(x, y) \ \mid J_{h^{-1}} \mid \tag{2.24}$$

où J_h désigne le Jacobien de la transformation h : $\quad J = \begin{vmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{vmatrix}$

Démonstration :

$$\begin{aligned}
\forall A \qquad \int\int_A f_{X,Y}(x, y) \ dx \ dy &= \int\int_{h(A)} f_{U,V}(u, v) \ du \ dv \\
&= \int\int_A f_{U,V}\Big(u(x, y), v(x, y)\Big) \ \mid J_h \mid \ dx \ dy
\end{aligned}$$

Cette égalité devant être vérifiée quel que soit A, il faut et suffit que les deux intégrandes soient égales.

2.5.3 Loi de probabilité d'une fonction de deux v.a. réelles

Le résultat précédent permet de calculer la densité de probabilité d'une fonction U du couple (X, Y), sous réserve que l'on puisse associer à U une autre fonction V du couple (X, Y) telle que la fonction qui à (X, Y) fait correspondre (U, V) soit bijective.
On applique le théorème précédent, puis on déduit la d.d.p. de la seule variable U en intégrant celle du couple (U, V) par rapport à V.

$$f_U(u) = \int_{-\infty}^{+\infty} f_{U,V}(u, v) \ dv$$

Exemple : détermination de la loi de $U = X + Y$. On peut associer à U la variable $V = X$, de sorte que les couples (X, Y) et (U, V) sont liés par une bijection h :

$h : U = X + Y \qquad h^{-1} : X = V \qquad J_h = J_{h^{-1}} = 1$
$\quad V = X \qquad\qquad Y = U - V$

$$f_U(u) = \int_{-\infty}^{+\infty} f_{U,V}(u,v)\ dv = \int_{-\infty}^{+\infty} f_{X,Y}(v, u-v)\ \mid 1 \mid\ dv$$

Notons le cas particulier important où X et Y sont indépendantes :

$$f_U(u) = f_X(u)\ *\ f_Y(u)$$

2.5.4 Loi du χ^2 à deux degrés de liberté

X et Y étant deux variables indépendantes gaussiennes centrées réduites, quelle est la loi de probabilité de la somme de leurs modules carrés ?
Ce problème se résout comme l'exemple précédent, à l'aide d'un changement de variables bijectif.
Soient $U = X^2 + Y^2$ et $V = \arctan(\frac{Y}{X})$.
Alors $X = \sqrt{U}\ \cos(V)$ et $Y = \sqrt{U}\ \sin(V)$, ce qui conduit à un jacobien égal à $\frac{1}{2}$. Donc

$$\begin{aligned} f_U(u) &= \int_0^{2\pi} f_{U,V}(u,v)\ dv \\ &= \int_0^{2\pi} f_{X,Y}(\sqrt{u}\cos v, \sqrt{u}\sin v)\ \frac{1}{2}\ dv \\ &= \int_0^{2\pi} \frac{1}{4\pi}\ e^{-\frac{u\cos^2 v + u sin^2 v}{2}}\ dv \\ &= \frac{1}{2}\ e^{-\frac{u}{2}} \qquad \text{pour } u \geq\ 0. \end{aligned}$$

2.6 Fonction caractéristique

La fonction exponentielle ayant un développement en série entière dont le rayon de convergence est infini, elle joue encore un rôle privilégié vis à vis des variables aléatoires. En effet :

$$E(e^{aX}) = E(\sum_{n=0}^{+\infty} \frac{a^n}{n!} X^n) \quad = \sum_{n=0}^{+\infty} \frac{a^n}{n!} E(X^n) \quad = \int_{-\infty}^{+\infty} e^{ax}\ f_X(x)\ dx \tag{2.25}$$

Remarquons que dans cette dernière intégrale, si on pose $a = iu$, on obtient la transformée de Fourier inverse de la densité de probabilité. Or, par définition, une densité de probabilité est une fonction positive sommable. Elle possède donc toujours une transformée de Fourier !

2.6.1 (Première) fonction caractéristique d'une v.a. réelle

2.6.1.1 Définition

Soient X une v.a. réelle et t un paramètre réel. On s'intéresse à la v.a. complexe $Y = \exp(itX)$.
La fonction φ_X de $\mathbb{R}$ dans $\mathbb{C}$, qui à t associe $E(Y)$ est la fonction caractéristique de la v.a. X.

$$\begin{aligned} \varphi_X(t) = &\int_{-\infty}^{+\infty} e^{itx}\ f_X(x)\ dx && \text{(v.a. continue)} \\ = &\sum_k e^{itx_k}\ \text{proba}(X = x_k) && \text{(v.a. discrète)} \end{aligned}$$

2.6.1.2 Propriétés

En tant que transformée de Fourier inverse, la fonction caractéristique est continue (cf Lebesgue) :

$$\varphi_X(t) = TF^{-1}\{f_X(\frac{t}{2\pi})\}$$

Elle est également bornée :

$$| \varphi_X(t) | \leq \int_{-\infty}^{+\infty} | e^{itx} \ f_X(x) | \ dx = \int_{-\infty}^{+\infty} f_X(x) \ dx = 1 = \varphi_X(0)$$

2.6.1.3 Développement de la fonction caractéristique

Théorème : si la v.a. X possède des moments jusqu'à l'ordre k, alors sa fonction caractéristique φ_X admet le développement limité

$$\varphi_X(t) = 1 + itE(X) + \cdots + \frac{(it)^k}{k!} E(X^k) + t^k \varepsilon(t) \qquad \text{avec} \qquad \varepsilon(0^+) = 0$$

Théorème : si la v.a. X possède des moments jusqu'à l'ordre k, alors sa fonction caractéristique φ_X est k fois dérivable et

$$\begin{aligned} \varphi_X^{(k)}(t) &= \int_{-\infty}^{+\infty} (ix)^k \ e^{itx} \ f_X(x) \ dx \\ \varphi_X^{(k)}(0) &= i^k \ E(X^k) \end{aligned}$$

En conséquence, le développement limité à l'ordre k de la fonction caractéristique autour de l'origine, s'il existe, permet de déterminer les moments d'ordre 1 à k de la variable aléatoire.

2.6.1.4 Addition de v.a. indépendantes

Soit $Z = X + Y$, où X et Y désignent deux v.a. indépendantes. On vérifie aisément $\varphi_Z(t) = E\{e^{it(X+Y)}\} = \varphi_X(t) \ \varphi_Y(t)$, soit, par transformée de Fourier inverse, $f_Z(z) = f_X(z) \ * \ f_Y(z)$.
Attention : la réciproque est toujours fausse ! $\varphi_Z(t) = \varphi_X(t) \ \varphi_Y(t) \quad \not\Rightarrow \quad X$ et Y indépendantes.

2.6.2 Fonction caractéristique d'une v.a. vectorielle

Soit $\underline{X} = (X_1, \ldots, X_n)$ une v.a. dans $\mathbb{R}^n$. Sa fonction caractéristique, $\varphi_{\underline{X}}$, est une fonction de $\mathbb{R}^n$ dans $\mathbb{C}$:

$$\varphi_{\underline{X}}(t1, \ldots, t_n) \triangleq E\{e^{i(X_1 t_1 + \cdots + X_n t_n)}\} \tag{2.26}$$

Elle possède les mêmes propriétés que dans le cas d'une variable réelle. En particulier, sous réserve d'inversibilité : $(2\pi)^n \ f_{\underline{X}}(\underline{x}) = \int \ldots \int_{\mathbb{R}^n} \ e^{-i(x_1 t_1 + \cdots + x_n t_n)} \ \varphi_{\underline{X}}(t_1, \ldots, t_n) \ dt_1 \ldots \ dt_n$
D'autre part, par marginalisation de la fonction vectorielle, on retrouve la fonction caractéristique d'une seule variable : $\varphi_{\underline{X}}(t_1, 0, \ldots, 0) = \varphi_{X_1}(t_1)$. Enfin, une condition nécessaire et suffisante d'indépendance de deux variables est que la fonction caractéristique du couple soit égale au produit des deux fonctions marginales : $\varphi_{X,Y}(u, v) = \varphi_X(u)\varphi_Y(v)$.

2.6.3 Seconde fonction caractéristique d'une v.a. réelle

Dans certaines disciplines, il arrive que l'on manipule le log de la fonction caractéristique. Ceci conduit à la **seconde fonction caractéristique** d'une v.a. : $\psi_X(t) = ln[\varphi_X(t)]$. L'intérêt majeur de cette définition est que la fonction caractéristique de la somme de deux v.a. indépendantes n'est autre que la somme des fonctions caractéristiques de chacune des variables.
Le développement limité de $\psi_X(t)$ autour de l'origine conduit aux **cumulants**, qui diffèrent des moments au delà de l'ordre deux.

2.7 Courbes de régression

Soient deux v.a. réelles X et Y qui admettent une densité de probabilité conjointe $f_{X,Y}$ et des densités marginales f_X et f_Y. Les courbes de régression ont pour objectif d'établir un éventuel lien fonctionnel entre ces deux variables aléatoires.

2.7.1 Problème

Quelle est la fonction g de X qui approche "au mieux" Y ? On cherche une fonction g telle que la différence $Y - g(X)$ soit la plus proche possible de zéro. L'égalité étant généralement impossible à assurer pour des variables aléatoires, on va s'intéresser à une erreur moyenne. Enfin, annuler une espérance mathématique peut s'avérer insuffisant, en raison de la dispersion des variables autour de leur moyenne. Pour y remédier, on aspire à minimiser l'erreur quadratique moyenne, c'est à dire que l'on cherche la fonction g qui minimise le critère $E\{[Y - g(X)]^2\}$.

$$\begin{aligned} E\{[Y-g(X)]^2\} &= \int\int_{\mathbb{R}^2} [y-g(x)]^2 \; f_{X,Y}(x,y) \; dx \; dy \\ &= \int_{\mathbb{R}} f_X(x) \int_{\mathbb{R}} \underbrace{[y-g(x)]^2 \; f_Y(y \mid x)}_{l(x,y,g(x))} \; dy \; dx \end{aligned}$$

La densité f_X et l'intégrande $l(x, y, g(x)) = [y - g(x)]^2 \; f_Y(y \mid x)$ étant toutes deux positives, il suffit de minimiser l'intégrale de la fonction l selon g. Pour ce faire, on annulle sa dérivée par rapport au nombre $g(x)$

$$\frac{\partial}{\partial(g(x))} \int_{\mathbb{R}} l() \; dy = \int_{\mathbb{R}} -2 \; [y - g(x)] \; f_Y(y \mid x) \; dy$$

Cette quantité est nulle si

$$\int_{\mathbb{R}} y \; f_Y(y \mid x) \; dy = \int_{\mathbb{R}} g(x) \; f_Y(y \mid x) \; dy = g(x) \int_{\mathbb{R}} f_Y(y \mid x) \; dy = g(x)$$

On obtient finalement $g(x) = m_Y(x) = \int_{\mathbb{R}} y \; f_Y(y \mid x) \; dy$.

2.7.2 Définition

Si elle existe, la quantité $m_Y(x)$ est appelée **moyenne conditionnelle** de Y par rapport à X. C'est une variable aléatoire, puisqu'elle est fonction de la v.a. X.
La courbe d'équation $y = m_Y(x)$ définit quant à elle la **courbe de régression** de Y par rapport à X.
De la même manière, il existe une régression $x = m_X(y)$ qui, dans le cas général, diffère de la précédente. Pour s'en convaincre, il suffit d'examiner les deux cas particuliers extrêmes :

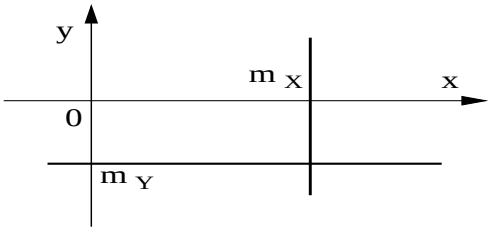

FIGURE 2.1 – X et Y indépendantes.

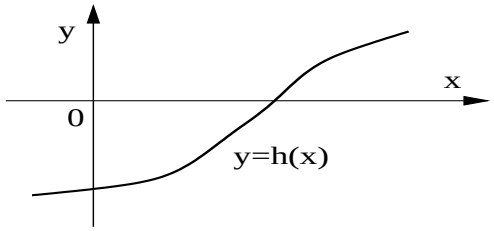

FIGURE 2.2 – bijection : $Y = h(X)$.

$$\begin{cases} m_Y(x) = E(Y) \\ m_X(y) = E(X) \end{cases} \qquad \begin{cases} m_Y(x) = h(x) \\ m_X(y) = h^{-1}(y) \end{cases}$$

2.7.3 Régression linéaire

Les résultats précédents nécessitent la connaissance de la densité de probabilité du couple (X, Y). A défaut, il est courant de se restreindre à une régression linéaire, soit une fonction g du type $g(X) = aX + b \approx Y$. Il faut alors déterminer les deux coefficients réels a et b qui minimisent l'erreur quadratique moyenne $E\{[Y - (aX + b)]^2\}$.

$$\begin{cases} \frac{\partial}{\partial a} & = -2E\{X(Y - aX - b)\} = 0 \\ \frac{\partial}{\partial b} & = -2E\{Y - aX - b\} = 0 \end{cases} \tag{2.27}$$

Soit

$$\begin{cases} E(XY) = aE(X^2) + bE(X) \\ E(Y) = aE(X) + b \end{cases} \tag{2.28}$$

Ce système linéaire se résout de façon classique

$$\begin{aligned} a_* &= \frac{Cov(X,Y)}{Var(X)} = r\ \frac{\sigma_Y}{\sigma_X} \\ b_* &= E(Y) - r\ \frac{\sigma_Y}{\sigma_X}\ E(X) \end{aligned} \tag{2.29}$$

L'approximation correspondante de Y s'écrit finalement

$$Y \approx g(X) = r\ \frac{\sigma_Y}{\sigma_X}\ [X - E(X)] + E(Y) \tag{2.30}$$

En permutant X et Y, on obtiendrait de la même manière

$$X \approx h(Y) = r\ \frac{\sigma_X}{\sigma_Y}\ [Y - E(Y)] + E(X) \tag{2.31}$$

On retrouve le fait que, si le coefficient de corrélation r est nul, les deux droites (2.30) et (2.31) sont parallèles aux axes, tandis qu'elles sont confondues si r est de module unitaire.
Attardons nous sur quelques propriétés remarquables du régresseur linéaire optimal :
– L'erreur d'approximation ($\tilde{Y} = Y - g(X)$) est centrée

$$\begin{aligned} E\Big\{\tilde{Y}\Big\} &= E\Big\{Y - a_*X - m_Y + a_*\ m_X\Big\} \\ &= E\Big\{Y - m_Y\Big\} - a_*\ E\Big\{X - m_X\Big\} \\ &= 0 \end{aligned} \tag{2.32}$$

Il s'agit en fait de la seconde équation du système (2.27).
– L'erreur d'approximation et le régresseur sont orthogonaux

$$\begin{aligned} E\Big\{\tilde{Y}X\Big\} &= E\Big\{(Y - a_*X - m_Y + a_*\ m_X)X\Big\} \\ &= E\Big\{(Y - m_Y)X\Big\} - a_*\ E\Big\{(X - m_X)X\Big\} \\ &= Cov(Y,X) - \frac{Cov(X,Y)}{Var(X)}\ Cov(X,X) \\ &= 0 \end{aligned} \tag{2.33}$$

C'est la conséquence de la première équation du système(2.27).

– L'erreur d'approximation est donc orthogonale à toute combinaison linéaire de 1 et X

$$E\Big\{\tilde{Y}(cX+d)\Big\} = c\ E\Big\{\tilde{Y}X\Big\} + d\ E\Big\{\tilde{Y}\Big\} = 0 \tag{2.34}$$

En particulier, l'erreur $\tilde{Y}$ et l'approximation optimale $g(X)$ sont orthogonales.

– L'erreur quadratique résultante n'est autre qu'une illustration du théorème de Pythagore

$$\begin{aligned} E\{Y^2\} &= E\Big\{\Big(g(X)+\tilde{Y}\Big)^2\Big\} \\ &= E\Big\{g^2(X)\Big\} + 2E\Big\{g(X)\ \tilde{Y}\Big\} + E\{\tilde{Y}^2\} \\ &= E\Big\{g^2(X)\Big\} + E\{\tilde{Y}^2\} \\ &= E\Big\{g^2(X)\Big\} + Var(\tilde{Y}) \end{aligned} \tag{2.35}$$

– On en déduit l'erreur quadratique optimale

$$\begin{aligned} Var(\tilde{Y}) &= E\Big\{Y^2\Big\} - E\Big\{g^2(X)\Big\} \\ &= E\Big\{Y^2\Big\} - E\Big\{(a_*X - a_*\ m_X + m_Y)^2\Big\} \\ &= E\Big\{Y^2\Big\} - a_*^2\ E\Big\{(X-m_X)^2\Big\} - 2a_*\ m_Y\ E\Big\{(X-m_X)\Big\} - m_Y^2 \\ &= Var(Y) - a_*^2\ Var(X) \\ &= Var(Y) - \frac{Cov^2(X,Y)}{Var(X)} \\ &= (1-r^2)\ \sigma_Y^2 \end{aligned} \tag{2.36}$$

Cette quantité positive devient nulle dès que le coefficient de corrélation est de module unitaire. Dans ce cas, cela signifie que, en moyenne quadratique, la v.a. centrée $Y - E(Y)$ est égale à $\frac{\sigma_Y}{\sigma_X}\ [X - E(X)]$.

2.7.4 Régression linéaire vectorielle

Lorsque les variables aléatoires X ou Y sont vectorielles, on peut établir une régression linéaire selon une formule analogue à la précédente.

$$\underline{Y} \approx \underline{\underline{a}}\ \underline{X} + \underline{b} \qquad \text{où } \underline{\underline{a}} \text{ et } \underline{b} \text{ minimisent le scalaire} \quad \varepsilon^2 = E\Big(\ ||\underline{\underline{a}}\ \underline{X} + \underline{b} - \underline{Y}||^2\ \Big)$$

Calculons les différentielles par rapport à chaque coefficient a_{ij} et b_i

$$\begin{cases} \frac{\partial \varepsilon^2}{\partial a_{ij}} = 2E\Big((\sum_k a_{ik}\ X_k + b_i - Y_i)\ X_j\Big) \\ \frac{\partial \varepsilon^2}{\partial b_i} = 2E\Big((\sum_k a_{ik}\ X_k + b_i - Y_i)\ 1\Big) \end{cases} \tag{2.37}$$

Annulons conjointement ces différentielles

$$\begin{cases} \underline{\underline{0}} &= E\Big([\underline{\underline{a}}\ \underline{X} + \underline{b} - \underline{Y}]\ \underline{X}'\Big) \\ \underline{0} &= E\Big([\underline{\underline{a}}\ \underline{X} + \underline{b} - \underline{Y}]\Big) \end{cases} \tag{2.38}$$

La première équation traduit l'orthogonalité de l'erreur minimale par rapport à la variable $\underline{X}$, tandis que la seconde établit la nullité de l'erreur moyenne.
Par linéarité de l'espérance mathématique

$$\begin{cases} \underline{\underline{0}} &= \underline{\underline{a}}E(\underline{X}\ \underline{X}') + \underline{b}E(\underline{X}') - E(\underline{Y}\ \underline{X}') \\ \underline{0} &= \underline{\underline{a}}E(\underline{X}) + \underline{b} - E(\underline{Y}) \end{cases} \tag{2.39}$$

$$\begin{cases} \underline{b} & = & E(\underline{Y}) - \underline{\underline{a}} E(\underline{X}) \\ \underline{\underline{a}} \underbrace{\Big(E(\underline{X}\ \underline{X}') - E(\underline{X})E(\underline{X}') \Big)}_{\underline{\underline{C}}_{XX}} & = & \underbrace{E(\underline{Y}\ \underline{X}') - E(\underline{Y})E(\underline{X}')}_{\underline{\underline{C}}_{YX}} \end{cases} \tag{2.40}$$

Lorsque la matrice $\underline{\underline{C}}_{XX}$ de variance covariance de la variable $\underline{X}$ est inversible

$$\begin{cases} \underline{\underline{a}} & = & \underline{\underline{C}}_{YX}\ \underline{\underline{C}}_{XX}^{-1} \\ \underline{b} & = & \underline{m}_Y - \underline{\underline{a}}\ \underline{m}_X \end{cases} \tag{2.41}$$

L'approximation résultante est analogue à celle établie dans le cas scalaire

$$\underline{Y} \approx \underline{\underline{C}}_{YX}\ \underline{\underline{C}}_{XX}^{-1} (\underline{X} - \underline{m}_X) + \underline{m}_Y \tag{2.42}$$

2.8 Théorèmes aux limites

La plupart des résultats que nous allons exploiter en statistiques s'appuie sur la convergence de suites de variables aléatoires. Mais que signifie l'égalité $\lim_{n\to+\infty} X_n = X$ lorsqu'elle implique des variables aléatoires ? Pour répondre à cette question, il faut d'abord répertorier les divers types d'existence de cette limite. Il en existe essentiellement quatre.

2.8.1 Divers modes de convergence

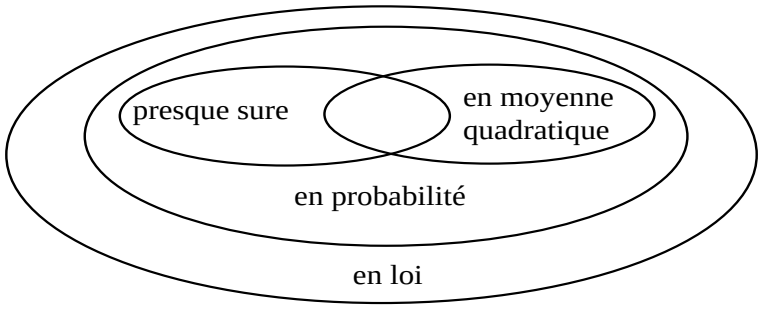

FIGURE 2.3 – les quatre modes de convergence.

- **convergence en loi** la suite des fonctions de répartition F_{X_n} converge vers la fonction de répartition F_X en tout point de continuité de F_X.
- **convergence en probabilité** $\forall h > 0 \qquad \lim_{n\to+\infty} \text{proba}(\mid X_n - X \mid > h) = 0.$
- **convergence en moyenne quadratique** $\lim_{n\to+\infty} E([X_n - X]^2) = 0.$
- **convergence presque sûre** $p(\ \{\omega \in \Omega / \lim_{n\to+\infty} X_n(\omega) = X(\omega)\}\ ) = 1.$

2.8.2 Lois des grands nombres

Considérons n v.a. X_i de même moyenne m et même écart-type σ. On en prend la moyenne empirique : $Z_n = \frac{1}{n} \sum_{i=1}^{n} X_i$, ce qui constitue une nouvelle variable aléatoire.

On sait que l'espérance mathématique de Z_n n'est autre que m, mais que peut-on dire de cette v.a. Z_n quand on fait tendre le nombre n vers l'infini ?

2.8.2.1 Loi faible

Les X_i sont deux à deux non corrélées. Calculons la variance de Z_n. Notons tout d'abord que la variable centrée peut s'écrire

$$Z_n - m = \frac{1}{n}\sum_{i=1}^{n} X_i \; - \; \frac{1}{n}\sum_{i=1}^{n} m = \frac{1}{n}\sum_{i=1}^{n} (X_i \; - \; m)$$

Alors

$$\begin{aligned} Var(Z_n) &= E\Big\{[\frac{1}{n}\sum_{i=1}^{n}(X_i \; - \; m)]^2\Big\} \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E\Big\{(X_i \; - \; m)(X_j \; - \; m)\Big\} \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \; + \; \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} Cov(X_i, X_j) \\ &= \frac{1}{n^2}\, n\, \sigma_X^2 \; + \; 0 \\ &= \frac{\sigma_X^2}{n} \end{aligned}$$

On en conclut que la variance de Z_n tend vers 0 quand n tend vers l'infini, c'est à dire que la suite des v.a. Z_n converge en moyenne quadratique vers le nombre m.

2.8.2.2 Loi forte

Les X_i sont deux à deux indépendantes. On montre alors que la suite des v.a. Z_n converge presque sûrement vers le nombre m.

2.8.3 Théorème de la limite centrale

On considère maintenant une suite de v.a. X_k, deux à deux indépendantes, identiquement distribuées (elles suivent toutes la même loi) et qui possèdent des moments au moins jusqu'à l'ordre deux. En particulier, elles ont donc même moyenne et même variance, finies.

On construit alors la nouvelle variable $W_n = \sum_{k=1}^{n} \frac{X_k - m}{\sigma\ \sqrt{n}}$, dont on montre qu'elle tend en loi vers une gaussienne centrée, réduite.

La démonstration s'appuie sur la fonction caractéristique de W_n qui, en vertu de l'indépendance des X_i, est donnée par le produit des fonctions caractéristiques des $\frac{X_k - m}{\sigma\sqrt{n}}$.

$$\varphi_{W_n}(t) = \prod_{k=1}^{n} E\{e^{it\frac{X_k - m}{\sigma\ \sqrt{n}}}\} = e^{-it\frac{m\ \sqrt{n}}{\sigma}}\ \{\varphi_X(\frac{t}{\sigma\ \sqrt{n}})\}^n$$

Or, φ_X admet un développement limité au moins jusqu'à l'ordre 2 puisque, par hypothèse, les deux premiers moments de X (m et $\sigma^2 + m^2$) existent.

$$\begin{aligned} \varphi_X(\frac{t}{\sigma\sqrt{n}}) &= 1 + iE(X)\ \{\ \frac{t}{\sigma\sqrt{n}}\ \} - \frac{1}{2}E(X^2)\ \{\ \frac{t}{\sigma\sqrt{n}}\ \}^2 + O(\ \{\ \frac{t}{\sigma\sqrt{n}}\}^3\ ) \\ &= 1 + \{\ \frac{imt}{\sigma\sqrt{n}}\ \} - \frac{(\sigma^2 + m^2)t^2}{2\sigma^2 n} + O(\ \{\ \frac{t}{\sigma\sqrt{n}}\}^3\ ) \end{aligned}$$

Alors, relativement à l'infiniment petit $\frac{t}{\sigma\sqrt{n}}$

$$\begin{aligned} ln\{\varphi_X(\frac{t}{\sigma\sqrt{n}})\} &= \{\ \frac{imt}{\sigma\sqrt{n}} - \frac{(\sigma^2+m^2)t^2}{2\sigma^2 n}\ \} - \frac{1}{2}\ \{\ \frac{imt}{\sigma\sqrt{n}} - \frac{(\sigma^2+m^2)t^2}{2\sigma^2 n}\ \}^2 + O(\ \{\ \frac{t}{\sigma\sqrt{n}}\}^3\ ) \\ &= \frac{imt}{\sigma\sqrt{n}} - \frac{\sigma^2 t^2}{2\sigma^2 n} + O(\ \{\ \frac{t}{\sigma\sqrt{n}}\}^3\ ) \end{aligned}$$

Ainsi

$$\begin{aligned} ln\{\varphi_{W_n}(t)\} &= \frac{-imt\sqrt{n}}{\sigma} + n\left\{\frac{imt}{\sigma\sqrt{n}} - \frac{t^2}{2n} + O(\ \{\ \frac{t}{\sigma\sqrt{n}}\}^3\ )\right\} \\ &= -\frac{t^2}{2} + \frac{1}{\sqrt{n}}\ O(\ \{\ \frac{t}{\sigma}\}^3\ ) \end{aligned}$$

Finalement

$$\lim_{n\to+\infty}\ \varphi_{W_n}(t) = \exp(-\frac{t^2}{2})$$

ce qui établit une convergence en loi.
Quelle que soit la loi des variables initiales X_k, pour peu qu'elles soient indépendantes et de variance finie, la fonction caractéristique de W_n tend vers celle d'une gaussienne centrée réduite.
Un contre-exemple classique résulte de la somme de deux variables indépendantes de loi de Cauchy. En calculant la fonction caractéristique d'une telle somme, on constate que cette dernière suit toujours une loi de Cauchy. Cette loi étant stable pour l'addition, elle ne tendra jamais vers une gaussienne. Mais ceci ne contredit en rien le théorème de la limite centrale, car la loi de Cauchy ne possède aucun moment !

2.9 Quelques lois continues usuelles

Dans tous les exemples qui suivent, la fonction échelon est représentée par $u(x)$, qui vaut 1 pour x positif et 0 pour x négatif.
La fonction Γ , de $\mathbb{R}^{*+}$ dans $\mathbb{R}^{*+}$, est définie par

$$\Gamma(p)\ =\ \int_0^{+\infty}\ e^{-x}\ x^{p-1}\ dx$$

La fonction beta, notée B, de $\mathbb{R}^{*+}\times\mathbb{R}^{*+}$ dans $\mathbb{R}^{*+}$, est définie par

$$B(p,q) = \frac{\Gamma(p)\ \Gamma(q)}{\Gamma(p+q)}\ =\ \int_0^{+\infty}\ \frac{x^{p-1}}{(1+x)^{p+q}}\ dx\ =\ \int_0^1\ x^{p-1}\ (1-x)^{q-1}\ dx$$

2.9.1 La loi de Gauss

2.9.1.1 Variable aléatoire réelle

X suit $\mathcal{N}(m,\sigma^2)$.
La densité de probabilité s'écrit $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{(x-m)^2}{2\sigma^2})$.
La fonction de répartition n'a pas d'expression analytique, mais elle est tabulée à l'aide de la fonction baptisée "Erf"

$$\begin{aligned} \mathrm{Erf}(x) &= \frac{2}{\sqrt{\pi}}\ \int_0^x\ e^{-u^2}\ du \\ F_X(x) &= \frac{1}{2}\ \left\{1 + \mathrm{Erf}(\frac{x-m}{\sigma\sqrt{2}})\right\} \end{aligned} \tag{2.43}$$

La fonction caractéristique possède elle aussi l'allure d'une gaussienne $\varphi_X(u) = e^{ium}\ e^{-\ \frac{\sigma^2\ u^2}{2}}$.
L'espérance mathématique est égale au paramètre m (cf équation(2.17)).
La variance est donnée par le terme σ^2 (cf équation(2.18)).
Tous les moments d'ordre supérieur à 2 se déduisent de ces deux paramètres.
La combinaison linéaire de deux gaussiennes indépendantes est encore une gaussienne. C'est évident dans le cas particulier de la somme de deux variables : $Z = X + Y$.

$$\varphi_Z(u) = \varphi_X(u)\ \varphi_Y(u) = e^{iu(m_X+m_Y)}\ e^{-\ \frac{u^2}{2}(\sigma_X^2\ +\ \sigma_Y^2)}$$

2.9.1.2 Variable aléatoire vectorielle

$\underline{X} = (X_1, X_2, \ldots, X_p)^t$.
La moyenne d'une telle variable est également vectorielle $\underline{m} = E(\underline{X}) = (E(X_1), E(X_2), \ldots, E(X_p))^t$.
Notons $\underline{X}_c$ la variable centrée $\underline{X}_c = \underline{X} - E(\underline{X}) = \underline{X} - \underline{m}$.
On définit sans difficulté la covariance d'une telle variable, en prenant garde au fait que cette quantité est une matrice, notée $\underline{\underline{C}}$ $\underline{\underline{C}} = E(\underline{X}_c\ \underline{X}_c^t)$.
En vertu de l'inégalité de Schwarz, cette matrice est définie positive (son inverse également par voie de conséquence), de déterminant positif c. Attention toutefois au cas particulier où l'une des variables dépend linéairement des autres car, alors la matrice de variance-covariance devient singulière et de rang $p-1$ (l'inégalité de Schwarz devient en effet une égalité en cas de dépendance linéaire totale).
Avec ces notations, l'expression de la densité de probabilité de la v.a. vectorielle centrée est analogue à celle d'une variable scalaire :

$$f_{\underline{X}}(\underline{x}) = \frac{1}{\sqrt{(2\pi)^p\ c}}\ \exp\{-\frac{1}{2}\ (\ \underline{x}_c^t\ \underline{\underline{C}}^{-1}\ \underline{x}_c\ )\ \} \tag{2.44}$$

Le cas particulier $p = 1$ conduit bien-sûr au cas scalaire exposé au paragraphe précédent.
La fonction caractéristique s'écrit

$$\varphi_{\underline{X}}(\underline{u}) = \exp(i\underline{u}^t\underline{m})\ \exp\Big\{-\frac{1}{2}(\underline{u}^t\ \underline{\underline{C}}\ \underline{u})\Big\} \tag{2.45}$$

Les lois marginales sont toutes gaussiennes.
Enfin, on justifie le cas particulier de la décorrélation de deux variables gaussiennes qui entraine leur indépendance. En effet, si les composantes X_i d'une gaussienne vectorielle sont décorrélées, la matrice de covariance correspondante $\underline{\underline{C}}$ est diagonale. La densité de probabilité de la v.a. vectorielle s'exprime immédiatement comme le produit des densités marginales. $f_{\underline{X}}(\underline{x}) = \prod_{i=1}^{p} f_{X_i}(x_i)$.
Ceci étant vrai quel que soit le point $\underline{x}$ considéré, on en déduit que les v.a. X_i sont indépendantes deux à deux.

2.9.2 La loi uniforme

Cette loi dépend de deux paramètres réels, a et $b > a$.
Densité de probabilité : $f_X(x) = \begin{array}{ll} \frac{1}{b-a} & \text{pour } a \le x \le b \\ 0 & \text{ailleurs} \end{array}$
Fonction de répartition : $F_X(x) = \begin{array}{ll} 0 & \text{pour } x < a \\ \frac{x-a}{b-a} & \text{pour } a \le x \le b \\ 1 & \text{pour } x > b \end{array}$
Fonction caractéristique : $\frac{e^{itb}-e^{ita}}{(b-a)\ it}$.

Moment d'ordre k : $E(X^k) = \frac{b^{k+1}-a^{k+1}}{(k+1)(b-a)}$.
On en déduit l'espérance mathématique $E(X) = \frac{b+a}{2}$
puis la variance $Var(X) = \frac{(b-a)^2}{12}$.

2.9.3 La loi log-normale

C'est la loi de la v.a. positive X donnée par $X = e^Y$, où Y suit $\mathcal{N}(m, \sigma^2)$.
Densité de probabilité : $f_X(x) = \dfrac{1}{x\sigma\sqrt{2\pi}}\ e^{-\frac{(\ln[x]-m)^2}{2\sigma^2}}\ u(x)$.
La fonction de répartition n'a pas d'expression analytique, mais elle est liée à celle de la gaussienne $F_X(x) = F_Y(ln[x]) = \mathrm{Erf}(ln[x])$.
Moment d'ordre k : $E(X^k) = e^{mk+\frac{k^2\sigma^2}{2}}$.
On en déduit l'espérance mathématique $E(X) = e^{m+\frac{\sigma^2}{2}}$
puis la variance $Var(X) = e^{2m+\sigma^2}\ (e^{\sigma^2} - 1)$.
Cette loi est fréquemment utilisée pour modéliser les durées de vie de composants électroniques, par exemple.

2.9.4 La loi gamma

Deux paramètres réels strictement positifs, p et θ.
Densité de probabilité : $f_X(x) = \frac{\theta^p}{\Gamma(p)}\ e^{-\theta x}\ x^{p-1}\ u(x)$.
Fonction caractéristique : $\varphi_X(t) = \frac{1}{(1-it/\theta)^p}$.
Moment d'ordre k : $E(X^k) = \frac{\Gamma(p+k)}{\theta^k\Gamma(p)}$
On en déduit la moyenne $E(X) = \frac{p}{\theta}$
et la variance $Var(X) = \frac{p}{\theta^2}$.

2.9.5 La loi exponentielle

Elle est un cas particulier de la précédente, obtenu pour $p = 1$.
Densité de probabilité : $f_X(x) = \theta\ e^{-\theta x}\ u(x)$.

2.9.6 La loi beta

C'est la loi du rapport de deux variables indépendantes, X et Y, suivant toutes deux une loi gamma, de paramètres respectifs $p > 0$ et $q > 0$.
Densité de probabilité : $f_Z(z) = \frac{1}{B(p,q)}\ \frac{z^{p-1}}{(1+z)^{p+q}}\ u(z)$.
Moment d'ordre k : $E(Z^k) = \frac{B(p+k,q-k)}{B(p,q)}$ pour $k < q$.
On en déduit la moyenne $E(Z) = \frac{p}{q-1}$ pour $q > 1$
et la variance $Var(Z) = \frac{p(p+q-1)}{(q-1)^2(q-2)}$ pour $q > 2$.

2.9.7 La loi de Laplace

Densité de probabilité : $f_X(x) = \frac{1}{2}\ e^{-|x|}$.
Fonction caractéristique : $\varphi_X(t) = \frac{1}{1+t^2}$.
Moment d'ordre k : $E(X^k) = 0$ k impair
$= k!$ k pair
On en déduit la moyenne $E(X) = 0$
et la variance $Var(X) = 2$.

2.9.8 La loi logistique

Densité de probabilité : $f_X(x) = \frac{e^{-x}}{(1+e^{-x})^2}$.
Fonction de répartition : $F_X(x) = \frac{1}{1+e^{-x}}$.
Moyenne $E(X) = 0$.
Variance : $\frac{\pi^2}{3}$.

2.9.9 La loi de Weibull

Deux paramètres réels strictement positifs : α et θ.
Densité : $f_X(x) = \alpha\theta x^{\alpha-1}\ e^{-\theta x^\alpha}\ u(x)$.
Fonction de répartition : $F_X(x) = \left(1 - e^{-\theta x^\alpha}\right) u(x)$.
Moyenne : $E(X) = \frac{\Gamma(1+1/\alpha)}{\theta^{1/\alpha}}$.
Variance : $Var(X) = \frac{\Gamma(1+2/\alpha) - \Gamma^2(1+1/\alpha)}{\theta^{2/\alpha}}$.

2.9.10 La loi du Khi-deux

C'est la loi de la somme des carrés de n variables aléatoires indépendantes suivant toutes une loi normale $\mathcal{N}(0,1)$: $U_n = \sum_{i=1}^{n} X_i^2$.
Le paramètre entier n désigne le nombre de degrés de liberté de la variable.
Densité : $f_{U_n}(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\ e^{-\frac{x}{2}}\ x^{\frac{n}{2}-1}\ u(x)$.
Moment d'ordre k : $E({U_n}^k) = 2^k \frac{\Gamma(\frac{n}{2}+k)}{\frac{n}{2}}$
On en déduit la moyenne $E(U_n) = n$
et la variance $Var(U_n) = 2n$.

2.9.11 La loi de Rayleigh

Soient X et Y deux v. a. gaussiennes, centrées, indépendantes et de même variance σ^2. La variable $Z = \sqrt{(X^2 + Y^2)}$ suit la loi de Rayleigh.
Densité : $f_Z(z) = \frac{z}{\sigma^2}\ e^{-\frac{z^2}{2\sigma^2}}\ u(z)$.
Moyenne : $E(Z) = \sigma\ \sqrt{\frac{\pi}{2}}$.
Variance : $Var(Z) = \sigma^2\ (2 - \frac{\pi}{2})$.

2.9.12 La loi de Student

C'est la loi du rapport entre une gaussienne centrée réduite X et la racine d'un Khi-deux à n degrés de liberté U_n, indépendante de X : $T_n = \frac{X}{\sqrt{\frac{U_n}{n}}}$.
Densité : $f_{T_n}(x) = \frac{1}{\sqrt{n}\ B(\frac{1}{2},\frac{n}{2})}\ \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$.
Moyenne : $E(T_n) = 0$.
Variance : $Var(T_n) = \frac{n}{n-2}$ pour $n > 2$.

2.9.13 La loi de Cauchy

C'est un cas particulier de la loi précédente, obtenu pour $n = 1$.
Densité : $f_{T_1}(x) = \frac{1}{\pi}\ \frac{1}{1+x^2}$.
Fonction caractéristique : $\varphi_{T_1}(t) = e^{-|t|}$.

Cette loi ne possède aucun moment !

2.9.14 La loi de Fisher-Snedecor

Soient X et Y deux variables aléatoires indépendantes suivant un Khi-deux à respectivement n et m degrés de liberté. La variable $U = \frac{X/n}{Y/m}$ suit une loi de Fisher-Snedecor à n et m degrés deliberté, notée $F_{n,m}$.
Densité : $f_U(x) = \frac{1}{B(\frac{n}{2},\frac{m}{2})}\ n^{\frac{n}{2}}\ m^{\frac{m}{2}}\ \frac{x^{n/2-1}}{(m+nx)^{\frac{n+m}{2}}}u(x)$.
Moyenne : $E(U) = \frac{m}{m-2}$, pour $m > 2$.
Variance : $Var(U) = \frac{2m^2\ (n+m-2)}{n\ (m-4)\ (m-2)^2}$, pour $m > 4$.

Chapitre 3

Statistiques

3.1 Introduction

L'objet des statistiques est l'étude d'une population constituée d'un grand nombre d'individus. Compte-tenu de ce grand nombre, on essaye de déterminer les caractéristiques "moyennes" de la population, sans examiner tous les individus séparément. On considère alors chaque élément observé comme une réalisation d'une variable aléatoire, dont la loi reflète l'ensemble de la population. Comment peut-on exploiter un grand nombre de données pour bien les synthétiser ? C'est le problème de la **statistique descriptive**, que nous n'aborderons pas ici.
La théorie des probabilités manipule essentiellement des variables aléatoires, dont on se donne la loi, de manière plus ou moins implicite. Les problèmes majeurs que rencontre le statisticien commencent à ce niveau : comment choisir la loi de probabilité qui correspond à un ensemble de mesures, considérées comme des réalisations d'une même variable aléatoire ?
Le type de loi lui-même peut découler du simple bon sens ou d'une grande expérience du statisticien, mais il reste généralement à ajuster un ou plusieurs coefficients caractérisant la loi. On se heurte alors aux problèmes d'**estimation de paramètres**.
Ensuite, il est bon de pouvoir s'assurer que la modélisation probabiliste retenue n'est pas trop éloignée de la réalité, ce qui nécessite de mettre en œuvre des **tests d'hypothèses**, pour valider les choix qui ont été faits.
Avant d'aborder ces deux points, il nous faut commencer par introduire un peu du vocabulaire propre aux statisticiens...

3.1.1 Echantillon

On s'intéresse à un phénomène qui affecte certains individus d'une population. Chaque mesure effectuée est considérée comme une réalisation de la variable aléatoire X qui modélise l'ensemble de la population.
Pour faire la moindre étude sur cette variable, on va prélever un ensemble de n individus qui, tant que la mesure n'est pas faite, est assimilable à une variable aléatoire vectorielle, dont chaque composante suit la même loi de probabilité, celle de X. On appelle cette variable **échantillon de taille n associé à X**. On la note généralement : $\mathcal{E} = \{X_1, \ldots, X_n\}$.
On dit de l'échantillon qu'il est **simple** quand les n variables X_i qui le constituent sont indépendantes.
On appelle **statistique** construite sur l'échantillon $\mathcal{E}$, toute fonction des n variables X_i. Il s'agit donc d'une nouvelle variable aléatoire.

L'exemple le plus simple est certainement la **moyenne empirique** qui permet d'estimer l'espérance mathématique d'une variable aléatoire : $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

On peut généraliser en considérant les moments empiriques : $\frac{1}{n}\sum_{i=1}^{n} X_i^k$.

D'où l'on déduit la **variance empirique** : $S_n^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2$.

Dans toute la suite du cours, le n-échantillon manipulé est supposé simple.

3.1.2 Théorème de Fisher

Un résultat revient souvent dans le cas gaussien, lorsque les v.a. X_i sont i.i.d. $\mathcal{N}(m, \sigma^2)$

1. $\overline{X}_n$ et S_n^2 sont indépendantes (surprenant !)
2. $\overline{X}_n$ suit $\mathcal{N}(m, \frac{\sigma^2}{n})$ (évident)
3. $n\,\frac{S_n^2}{\sigma^2}$ suit χ^2_{n-1} (presque évident)
4. $\frac{\overline{X}_n - m}{S_n}\sqrt{n-1}$ suit Student$(n-1)$ (exemple de changement de variables !)

Sa démonstration constitue un ensemble de quatre excellents exercices. Elle s'appuie essentiellement sur la décomposition itérative de la variance empirique d'un n-échantillon simple

$$nS_n^2 = \sum_{i=2}^{n} \frac{i-1}{i}\,(X_i - \overline{X}_{i-1})^2$$

On montre que les v.a. $Y_i = X_i - \overline{X}_{i-1}$ sont deux à deux décorrélées, ce qui, sous hypothèse gaussienne, entraîne l'indépendance. La suite en découle ...

3.2 Estimation de paramètres

Soit X une variable aléatoire de loi connue, à l'exception d'un ou plusieurs paramètres, noté(s) Φ, que l'on veut pouvoir estimer au vu d'un n-échantillon simple.
Il s'agit par exemple d'ajuster le paramètre réel positif λ d'une loi de Poisson. Rappelons au passage qu'une telle v.a. est discrète, à valeurs dans $\mathbb{N}$: proba$(X = x) = \frac{\lambda^x}{x!}\,e^{-\lambda}$.
Comment procéder et quelle confiance peut-on accorder au résultat ?
Les points qui suivent répondent partiellement à cette question.

3.2.1 Estimateur

Un estimateur exploite les n valeurs de l'échantillon dont on dispose. C'est généralement une fonction des n variables aléatoires X_i. Il s'agit donc d'une nouvelle variable aléatoire, notée $\hat{\Phi}_n$, qui est une statistique particulière construite sur le n-échantillon $\{X_1, X_2, \ldots, X_n\}$.
A titre d'exemple, on a déjà vu la moyenne empirique qui permet d'estimer le paramètre m d'une gaussienne : $\hat{\Phi}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

3.2.2 Consistance

Une propriété demandée à un estimateur est qu'il s'approche d'autant plus du paramètre inconnu que la taille n de l'échantillon augmente. Un estimateur est alors qualifié de consistant s'il converge en probabilité vers le paramètre recherché :

$$\forall \varepsilon > 0 \qquad \lim_{n \to +\infty} \text{proba}\Big\{ \mid \hat{\Phi}_n - \Phi \mid > \varepsilon \Big\} = 0$$

3.2.3 Biais

Un estimateur étant généralement une variable aléatoire destinée à approcher un nombre, il parait important de comparer la moyenne de cette v.a. au nombre en question. La différence entre ces deux termes définit ce que l'on appelle le **biais** de l'estimateur. Cette quantité peut dépendre de la taille n de l'échantillon utilisé, aussi est-elle indicée selon n.

$$b_n(\Phi) \triangleq E(\hat{\Phi}_n) - \Phi \tag{3.1}$$

On dit d'un estimateur qu'il est sans biais, ou non biaisé, si, pour tout n, $\quad b_n(\Phi) = 0$.
A défaut, un estimateur est asymptotiquement sans biais si $\lim_{n \to +\infty} b_n(\Phi) = 0$.

3.2.4 Calcul intermédiaire

Un estimateur sans biais peut s'avérer décevant si la dispersion de ses réalisations est importante vis à vis du nombre à évaluer. Il convient donc de quantifier cette dispersion en s'intéressant à la variance de l'estimateur. Dans ce but, la plupart des calculs qui vont suivre s'appuient sur un résultat intermédiaire que nous allons établir maintenant.

$$\frac{\partial^k}{\partial \Phi^k} E\{h(\underline{X})\} = \frac{\partial^k}{\partial \Phi^k} \int_{\mathbb{R}} h(\underline{x})\ f_{\underline{X}}(\underline{x})\ d\underline{x} \ = \ \int_{\mathbb{R}} h(\underline{x})\ \frac{\partial^k}{\partial \Phi^k}\ f_{\underline{X}}(\underline{x})\ d\underline{x} \tag{3.2}$$

Il convient de noter que cette formule suppose que la densité de probabilité $f_{\underline{X}}$ est dérivable sur $\mathbb{R}$, ce qui signifie que le domaine des réalisations des variables aléatoires X_i n'est pas lui-même fonction du paramètre à estimer Φ. (un contre-exemple flagrant est la loi uniforme dont on voudrait estimer une borne, voire les deux).
Deux valeurs particulières de k vont s'avérer utiles.
$k = 1$:

$$\begin{aligned}
\frac{\partial}{\partial \Phi} E\{h(\underline{X})\} &= \int_{\mathbb{R}} h(\underline{x})\ \frac{f_{\underline{X}}(\underline{x})}{f_{\underline{X}}(\underline{x})}\ \frac{\partial}{\partial \Phi}\ f_{\underline{X}}(\underline{x})\ d\underline{x} \\
&= \int_{\mathbb{R}} h(\underline{x})\ f_{\underline{X}}(\underline{x})\ \frac{\partial}{\partial \Phi}\ \ln f_{\underline{X}}(\underline{x})\ d\underline{x} \\
&= E\Big\{ h(\underline{X})\ \frac{\partial}{\partial \Phi}\ \ln[f_{\underline{X}}(\underline{X})] \Big\}
\end{aligned} \tag{3.3}$$

$k = 2$:

$$\begin{aligned}
\frac{\partial^2}{\partial \Phi^2} E\{h(\underline{X})\} &= \int_{\mathbb{R}} h(\underline{x})\ \frac{\partial}{\partial \Phi} \Big\{ \frac{f_{\underline{X}}(\underline{x})}{f_{\underline{X}}(\underline{x})}\ \frac{\partial}{\partial \Phi}\ f_{\underline{X}}(\underline{x}) \Big\}\ d\underline{x} \\
&= \int_{\mathbb{R}} h(\underline{x})\ \frac{\partial}{\partial \Phi} \Big\{ f_{\underline{X}}(\underline{x})\ \frac{\partial}{\partial \Phi}\ \ln[f_{\underline{X}}(\underline{x})] \Big\}\ d\underline{x} \\
&= \int_{\mathbb{R}} h(\underline{x}) \Big\{ \frac{\partial}{\partial \Phi}\ f_{\underline{X}}(\underline{x})\ \frac{\partial}{\partial \Phi}\ \ln[f_{\underline{X}}(\underline{x})] \ + \ f_{\underline{X}}(\underline{x})\ \frac{\partial^2}{\partial \Phi^2}\ \ln[f_{\underline{X}}(\underline{x})] \Big\}\ d\underline{x} \\
&= \int_{\mathbb{R}} h(\underline{x}) \Big\{ \ f_{\underline{X}}(\underline{x})\ (\frac{\partial}{\partial \Phi}\ \ln[f_{\underline{X}}(\underline{x})])^2 \ + \ f_{\underline{X}}(\underline{x})\ \frac{\partial^2}{\partial \Phi^2}\ \ln[f_{\underline{X}}(\underline{x})] \Big\}\ d\underline{x}
\end{aligned}$$

$$= E\Big\{h(\underline{X})\ (\frac{\partial}{\partial\Phi}\ \ln[f_{\underline{X}}(\underline{X})])^2\Big\}\ +\ E\Big\{h(\underline{X})\ \frac{\partial^2}{\partial\Phi^2}\ \ln[f_{\underline{X}}(\underline{X})]\Big\} \tag{3.4}$$

3.2.5 Vraisemblance

En statistiques, il est une fonction qui joue un rôle privilégié. Il s'agit de la fonction de vraisemblance, notée L, qui à la variable $(x_1, x_2, \ldots, x_n, \Phi)$ associe le nombre égal à $f_{\underline{X}}(x_1, x_2, \ldots, x_n)$. Tant que l'on considère, comme c'est le cas ici, que l'échantillon manipulé est simple, cette dernière quantité est donnée par le produit des densités marginales :

$$L(x_1, x_2, \ldots, x_n, \Phi) \triangleq \prod_{k=1}^{n}\ f_X(x_k) \tag{3.5}$$

3.2.6 Suffisance (Neyman et Fisher)

Une statistique $g(\underline{X})$ est qualifiée de suffisante si, et seulement si, il est possible de factoriser la fonction de vraisemblance L de l'échantillon sous la forme :

$$L(x_1, \ldots, x_n, \Phi) = \alpha\{g(x_1, \ldots, x_n), \Phi\}\ .\ \beta(x_1, \ldots, x_n) \tag{3.6}$$

où α et β sont deux fonctions non négatives qui ne dépendent que de leurs arguments.
Cette factorisation signifie que, seule la valeur de la statistique $g(\underline{x})$ amène de l'information sur le paramètre inconnu Φ, sans qu'il soit utile de savoir comment se répartissent les divers x_i antécédents de ce nombre $g(\underline{x})$. Lorsqu'elle est possible, elle n'est pas unique.

3.2.7 Quantité d'information (au sens de Fisher)

Selon Fisher, l'information relative au paramètre inconnu Φ, contenue dans le n-échantillon $\{X_1, \ldots, X_n\}$ est donnée par l'espérance mathématique :

$$I_n(\Phi) \triangleq E\Big\{\big[\frac{\partial}{\partial\Phi}\ln L(X_1, \ldots, X_n, \Phi)\big]^2\Big\} \tag{3.7}$$

Théorème : lorsque l'espace des réalisations de la variable aléatoire X ne dépend pas de Φ, cette quantité d'information, proportionnelle à la taille n de l'échantillon, n'est autre que la variance de la variable aléatoire $\frac{\partial \ln L}{\partial \Phi}$:

$$\begin{aligned} I_n(\Phi) &= \mathrm{Var}(\frac{\partial \ln L}{\partial \Phi}) \\ &= -E\Big\{\frac{\partial^2}{\partial\Phi^2}\ln L(X_1, \ldots, X_n, \Phi)\Big\} \\ &= nI_1(\Phi) \end{aligned} \tag{3.8}$$

La démonstration se fait en deux temps.
– il faut d'abord montrer que la v.a. $\frac{\partial \ln L}{\partial \Phi}$ est de moyenne nulle. pour cela, il est commode d'utiliser l'équation (3.3), dans laquelle on fait $h(\underline{X}) = 1$ et $k = 1$:

$$\frac{\partial}{\partial\Phi}E\{1\} = 0 = E\Big\{\ \frac{\partial}{\partial\Phi}\ \ln[f_{\underline{X}}(\underline{X})\Big\} \tag{3.9}$$

– il reste à calculer le moment d'ordre 2 de cette variable aléatoire. on considère maintenant l'équation (3.4), dans laquelle on fait toujours $h(\underline{X}) = 1$ mais $k = 2$:

$$\frac{\partial^2}{\partial\Phi^2}E\{1\} = 0 = E\Big\{(\frac{\partial}{\partial\Phi}\ \ln[f_{\underline{X}}(\underline{X})])^2\Big\}\ +\ E\Big\{\frac{\partial^2}{\partial\Phi^2}\ \ln[f_{\underline{X}}(\underline{X})]\Big\} \tag{3.10}$$

3.2.8 Inégalité de Cramer-Rao

Théorème : lorsque l'espace des réalisations de la variable aléatoire X ne dépend pas de Φ :

$$\mathrm{Var}(\hat{\Phi}_n) \geq \frac{\left(\frac{\partial}{\partial\Phi}E(\hat{\Phi}_n)\right)^2}{I_n(\Phi)} = \frac{\left(1+\dot{b}_n(\Phi)\right)^2}{I_n(\Phi)} \tag{3.11}$$

La démonstration fait à nouveau appel à l'équation (3.2), dans laquelle on pose $h(\underline{X}) = g(\underline{X}) = \hat{\Phi}_n$ et $k = 1$:

$$\begin{aligned}\frac{\partial}{\partial\Phi}E\{g(\underline{X})\} &= 1+\dot{b}_n(\Phi) \\ &= \int_{\mathbb{R}} g(\underline{x})\ \frac{\partial}{\partial\Phi}\ f_{\underline{X}}(\underline{x})\ d\underline{x} \\ &= E\Big\{g(\underline{X})\ \frac{\partial}{\partial\Phi}\ \ln[f_{\underline{X}}(\underline{X})]\Big\}\end{aligned} \tag{3.12}$$

Le dernier terme est le moment croisé d'ordre 2 de deux variables aléatoires, dont l'une au moins est de moyenne nulle. On peut donc retrancher le produit des moyennes, pour faire apparaitre la covariance de ces deux variables :

$$\begin{aligned}1+\dot{b}_n(\Phi) &= E\Big\{g(\underline{X})\ \frac{\partial}{\partial\Phi}\ \ln[f_{\underline{X}}(\underline{X})]\Big\} - E\Big\{g(\underline{X})\Big\}\ E\Big\{\frac{\partial}{\partial\Phi}\ \ln[f_{\underline{X}}(\underline{X})]\Big\} \\ &= \mathrm{Cov}\Big(g(\underline{X}), \frac{\partial}{\partial\Phi}\ \ln[f_{\underline{X}}(\underline{X})]\Big)\end{aligned} \tag{3.13}$$

Or, on a vu dans le cours de probabilités que le module carré d'une covariance est majoré par le produit des variances. Dans le cas présent, la traduction de cette inégalité conduit au théorème de Cramer-Rao :

$$|\ 1+\dot{b}_n(\Phi)\ |^2 \leq\ \mathrm{Var}\Big(g(\underline{X})\Big)\ \mathrm{Var}\Big(\frac{\partial}{\partial\Phi}\ \ln[f_{\underline{X}}(\underline{X})]\Big) \tag{3.14}$$

3.2.9 Efficacité

L'efficacité d'un estimateur est définie par le rapport

$$\mathrm{Eff}(\hat{\Phi}_n) \triangleq \frac{\left(\frac{\partial}{\partial\Phi}E(\hat{\Phi}_n)\right)^2}{I_n(\Phi)\ \mathrm{Var}(\hat{\Phi}_n)} \qquad \in\ [0,1] \tag{3.15}$$

Cette quantité étant naturellement comprise entre 0 et 1, on dit que l'estimateur $\hat{\Phi}_n$ est efficace lorsqu'elle vaut 1. L'estimateur est qualifié d'asymptotiquement efficace si $\lim\limits_{n\to+\infty} \mathrm{Eff}(\hat{\Phi}_n) = 1$.

3.2.10 Estimateur du maximum de vraisemblance

Comme son nom l'indique, cet estimateur maximise la vraisemblance. La fonction ln étant strictement croissante, on peut le déterminer en maximisant non pas la vraisemblance elle même, mais son ln, afin de simplifier un peu les calculs. $\hat{\Phi}_{MV}$ est la valeur de Φ qui satisfait l'équation

$$\frac{\partial}{\partial\Phi}\ln L(X_1,\ldots,X_n,\Phi) = 0 \tag{3.16}$$

Cet estimateur peut également s'interpréter comme la valeur du paramètre inconnu Φ qui minimise la "distance" entre la loi empirique de l'échantillon $(g_{\underline{X}})$ et la densité supposée $(f_{\underline{X}})$ paramétrée par Φ, c'est à dire la fonction de vraisemblance (L). La pseudo-distance que l'on minimise alors est celle de **Kullback-Leibler** :

$$K(g_{\underline{X}}, f_{\underline{X}}) \ \triangleq\ \int_{\mathbb{R}} g_{\underline{X}}\ (\underline{x})\ \ln\ \frac{g_{\underline{X}}\ (\underline{x})}{f_{\underline{X}}\ (\underline{x})}\ d\underline{x}$$

$$= \int_{\mathbb{R}} g_{\underline{X}}(\underline{x}) \ln g_{\underline{X}}(\underline{x}) \, d\underline{x} - \int_{\mathbb{R}} g_{\underline{X}}(\underline{x}) \ln f_{\underline{X}}(\underline{x}) \, d\underline{x} \qquad (3.17)$$

Il s'agit d'une pseudo-distance car elle n'est pas symétrique. Cependant, on montre que cette quantité positive n'est nulle que lorsque la densité $f_{\underline{X}}$ est égale, presque partout, à la densité $g_{\underline{X}}$. Minimiser cette distance revient donc à maximiser le dernier terme selon Φ

$$\int_{\mathbb{R}} g_{\underline{X}}(\underline{x}) \ln L(\underline{x}, \Phi) \, d\underline{x}$$

Ce résultat s'obtient en particulier pour la valeur du paramètre Φ solution de l'équation (3.16).

Exemple : estimation du paramètre λ d'une loi de Poisson : proba$(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$, $x \in \mathbb{N}$.
La fonction de vraisemblance d'un n-échantillon simple construit sur une telle loi vaut :

$$L(X_1, \ldots, X_n, \lambda) = \prod_{i=1}^{n} p(X = X_i) = \prod_{i=1}^{n} \frac{\lambda^{X_i}}{X_i!} e^{-\lambda} = e^{-n\lambda} \prod_{i=1}^{n} \frac{\lambda^{X_i}}{X_i!}$$

On en prend le ln :

$$\ln L(X_1, \ldots, X_n, \lambda) = -n\lambda + \sum_{i=1}^{n} \{X_i \ln(\lambda) - \ln(X_i!)\}$$

On différencie par rapport à λ :

$$\frac{\partial}{\partial \lambda} \ln L(\ldots) = -n + \sum_{i=1}^{n} \{X_i \frac{1}{\lambda}\} = -n + \frac{1}{\lambda} \sum_{i=1}^{n} X_i$$

L'estimateur au sens du maximum de vraisemblance est donc donné par :

$$\hat{\lambda}_{MV} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Dans ce cas particulier, il s'agit tout simplement de la moyenne empirique de l'échantillon.
A posteriori, on constate aisément que la statistique $\overline{X}$ est bel et bien suffisante

$$L(X_1, \ldots, X_n, \lambda) = e^{-n\lambda} \prod_{i=1}^{n} \frac{\lambda^{X_i}}{X_i!} = \left(e^{-n\lambda} \lambda^{n \overline{X}}\right) \left(\frac{1}{\prod_{i=1}^{n} X_i!}\right)$$

Examinons rapidement les qualités de cet estimateur :

- $E\{\hat{\lambda}_{MV}\} = \frac{1}{n} \sum_{i=1}^{n} E\{X_i\} = E\{X\} = \cdots = \lambda$ il est non biaisé
- $\mathrm{Var}\{\hat{\lambda}_{MV}\} = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{1}{n}\mathrm{Var}(X) = \cdots = \frac{\lambda}{n}$ il est consistant
  (ce calcul exploite l'indépendance des X_i : la variance de la somme est égale à la somme des variances).
- $\frac{\partial^2}{\partial \lambda^2} \ln L(\ldots) = -\frac{1}{\lambda^2} \sum_{i=1}^{n} X_i \qquad \Rightarrow \qquad I_n(\Phi) = -E\Big\{\frac{\partial^2}{\partial \Phi^2} \ln[L(\ldots)]\Big\} = \frac{n}{\lambda}$

On en conclut que cet estimateur non biaisé est efficace.

3.2.11 Intervalle de confiance

Un estimateur est une variable aléatoire : à chaque réalisation de l'échantillon $(x_1, \ldots, x_n)$ correspond une valeur $\hat{\phi}_n$. La question qui se pose est de savoir quelle confiance on peut accorder à une réalisation particulière $\hat{\phi}_n$. Pour répondre à cette interrogation, il faut d'abord déterminer la loi de la variable aléatoire $\hat{\Phi}_n$. Ensuite, il s'agit de définir un intervalle $[a_\alpha, b_\alpha]$ tel que, conformément à cette loi, la probabilité que cet intervalle recouvre effectivement le paramètre réel Φ soit égale à $100\alpha\%$. Il n'y a pas toujours unicité quant au choix des bornes d'un tel intervalle. On retient alors le plus étroit, parmi toutes les solutions possibles.

3.3 Tests d'hypothèses

Tous les paragraphes précédents visent à estimer un ou plusieurs paramètres de sorte que la loi attribuée a priori aux données "colle" au mieux avec l'échantillon obtenu.
Il est nécessaire de disposer d'outils permettant par ailleurs de quantifier l'adéquation entre les données et le modèle qu'elles sont censées suivre. Ce type de problème rentre dans le cadre général et difficile des tests d'hypothèses.
Il n'est pas question ici de procéder à une étude exhaustive de ce domaine, dans lequel la littérature est particulièrement abondante. Nous allons nous contenter de traiter quelques exemples simples, afin de dégager les principes de ces tests.
Pour commencer, l'ensemble des tests se décompose en plusieurs sous-ensembles, selon la nature des hypothèses à tester. D'un côté, on trouve les tests paramétriques, relatifs à la valeur d'un ou plusieurs paramètres dont dépend la loi de l'échantillon considéré. De l'autre figurent les tests non paramétriques, où la question envisagée est l'adéquation d'un modèle à l'échantillon.
Quel que soit le problème considéré, il faut bien comprendre que la conclusion sera tirée au vu d'un échantillon, réalisation d'une variable aléatoire vectorielle et que, en conséquence, il n'existe pas de réponse à 100%. On risque toujours de se tromper et tout l'art du test est de réduire autant que possible les erreurs.

3.3.1 Le test du Khi deux

Il s'agit d'un critère de conformité, qui vise à dire si un échantillon est issu ou non d'une répartition F_X donnée. Il existe plusieurs tests de ce type, mais le plus courant est sans doute le "test du Khi deux" :
- on fait une partition en l intervalles Δ_i de l'ensemble Ω des réalisations de la v.a. X.
- on comptabilise le nombre ν_i d'éléments x_k du n-échantillon qui appartiennent à chaque intervalle Δ_i.
- on calcule la probabilité théorique p_i qu'a la variable X de prendre ses valeurs dans Δ_i.
- on compare l'expérimental au théorique en construisant une nouvelle v.a. qui mesure une erreur quadratique :

$$\mathcal{C} = \sum_{i=1}^{l} \frac{1}{np_i} (\nu_i - np_i)^2 \tag{3.18}$$

Pearson a montré que si l'hypothèse sur la répartition F_X des X_k est vraie, alors la v. a. $\mathcal{C}$ tend asymptotiquement vers la loi du χ^2 à $l-1$ degrés de liberté et ce, indépendamment de la forme de la répartition F_X.
Malheureusement, on ne peut rien dire de particulier si l'hypothèse est fausse...
Exemple : Une machine automatisée fabrique des pièces à cadence élevée. Chaque pièce est soit mauvaise (probabilité p), soit bonne (probabilité $1-p$).

On prélève 100 échantillons (supposés simples) de 300 pièces chacun et on obtient les résultats suivants :

nb de pièces défectueuses par échantillon	0	1	2	3	4	5	6	7	8	≥ 9
nb d'échantillons	6	16	20	23	16	10	4	3	2	0

Si les inégalités de production sont dues au hasard, la variable aléatoire dont on vient d'acquérir un n-échantillon suit une loi donnée, que l'on peut tester.
On peut associer à chaque pièce une v.a. $V_{i,j}$ qui vaut 1 si la pièce est défectueuse (probabilité p), 0 sinon (probabilité $1-p$). L'indice i repère l'échantillon ($i = 1$ *à* 100), tandis que l'indice j désigne la pièce ($j = 1$ *à* 300). Le paramètre inconnu p s'estime directement comme la moyenne empirique des $V_{i,j}$:

$$\hat{p} = \frac{1}{100}\,\frac{1}{300}\sum_{i=1}^{100}\sum_{j=1}^{300} V_{i,j} = \cdots = 0,01$$

Les données représentent 100 réalisations, supposées indépendantes, de la variable aléatoire X égale à la somme sur j des $V_{i,j}$. Si les hypothèses relatives au hasard sont justes, cette v.a. suit une loi binômiale de paramètres $p = 0,01$ et $n = 300$. La moyenne de X vaut donc $m = np = 3$ tandis que sa variance est égale à $np(1-p) = 2,97$.
Pour tester cette hypothèse, il suffit maintenant de comparer l'histogramme expérimental à l'histogramme théorique. Cependant, les valeurs théoriques de la binômiale étant difficilement accessibles d'un point de vue numérique, on peut les approcher par une gaussienne ($m = 3$ et $\sigma^2 = 2,97$), ou par une loi de Poisson ($\lambda = np = 3$), puisque le paramètre p est "faible". Les calculs qui suivent résultent de l'approximation selon la loi de Poisson.

x	0	1	2	3	4	5	6	7	8	≥ 9
classe k	1	2	3	4	5	6	7	8	9	10
expérimental y_k	6	16	20	23	16	10	4	3	2	0
théorique p_k	4,98	14,94	22,40	22,40	16,80	10,08	5,04	2,16	0,81	r

Le nombre r qui figure en dernière position dans le tableau précédent dépend linéairement des autres valeurs numériques théoriques : $r = 100\ (1 - p_1 - \cdots - p_9) = 0,39$.
De plus, le paramètre p a été estimé à partir des données sur lesquelles on fait le test.
Partant d'un total de 10 classes, il convient donc de retrancher deux degrés de liberté. La variable aléatoire Z qui mesure l'écart-quadratique entre la pratique et la théorie supposée suit donc une loi du Khi deux à 8 degrés de liberté. Ici, elle prend comme réalisation :

$$z = \sum_{k=1}^{10} \frac{(y_k - 100\ p_k)^2}{100\ p_k} = \cdots = 3,35$$

La table du Khi-deux à 8 degrés de liberté donne [0 , 15.5] comme intervalle de confiance à 95%. Cet intervalle contient la valeur trouvée pour z : on accepte donc l'hypothèse.

3.3.2 Les tests paramétriques

L'exemple précédent illustre assez bien le genre de décisions qu'un statisticien essaie de prendre au vu d'un échantillon aléatoire. En général, il s'agit d'accepter ou de réfuter une hypothèse, éventuellement choisir entre deux hypothèses. Certaines méthodes introduisent une alternative supplémentaire en décrétant que l'on ne dispose pas de suffisamment d'information pour trancher, mais nous ne les traiterons pas ici.
Un test confronte deux hypothèses, traditionnellement notées H_0 et H_1. Celles ci peuvent être complémentaires, par exemple H_0 = *"la moyenne de l'échantillon est négative ou nulle"*, contre H_1 = *"la moyenne de l'échantillon est strictement positive"*. Mais elles peuvent n'être qu'incompatibles, par exemple H_0 = *"la moyenne de l'échantillon vaut -1"*, contre H_1 = *"la moyenne de l'échantillon est positive"*. On devine alors que les deux hypothèses ne jouent pas toujours un rôle symétrique. En fait, on privilégie généralement H_0 en majorant le risque de rejeter à tort cette hypothèse. Sous cette contrainte, on fait alors au mieux vis à vis de l'alternative H_1.
L'objectif d'un test consiste à définir une partition (G_0, G_1) de l'ensemble $\mathbb{R}^n$ des réalisations de l'échantillon telle que, si $\underline{x} \in G_0$ on accepte H_0 (décision γ_0), tandis que si $\underline{x} \in G_1$ on rejette H_0 (décision γ_1). Ce principe s'applique parfois plus simplement sur une statistique, soit une fonction de l'échantillon, plutôt que sur l'échantillon lui-même.
Le test du Khi-deux que nous venons de voir conduit à l'acceptation de l'hypothèse H_0 (l'échantillon est de loi binômiale) ou à l'hypothèse complémentaire $H_1 = \overline{H}_0$. C'est un exemple de test non paramétrique, en ce sens qu'il ne porte pas sur la valeur d'un ou plusieurs paramètres d'une loi donnée.
Les tests paramétriques sont relativement nombreux, chacun dépendant du cas particulier traité. Il n'est pas question de les passer tous en revue dans le cadre de ce cours, mais nous allons en examiner un exemple simple. Nous verrons ensuite le principe d'une autre méthode, fondée sur la minimisation d'un risque.

3.3.3 Un exemple de test paramétrique

Etudions la méthode classique au travers d'un petit exemple, visant à déterminer si la moyenne inconnue m d'un n-échantillon gaussien $\underline{X}$ de variance connue σ^2 est négative ou nulle (hypothèse H_0) ou bien si cette moyenne est positive (hypothèse H_1).
Quatre cas de figure se présentent, auxquels on attribue conventionnellement les probabilités :
- on décide γ_0 et H_0 est vraie : probabilité $1 - \alpha$ = **niveau de confiance** du test.
- on décide γ_1 et H_0 est vraie : probabilité α = **seuil** ou **niveau de signification** du test (c'est l'erreur de $1^{ère}$ espèce, à éviter en priorité).
- on décide γ_0 et H_1 est vraie : probabilité $1 - \beta$ (c'est l'erreur de 2^{nde} espèce).
- on décide γ_1 et H_1 est vraie : probabilité β = **puissance** du test.

décision réalité	γ_0	γ_1
H_0	$1-\alpha$	α
H_1	$1-\beta$	β

Par hypothèse, la variable aléatoire $\underline{X}$ suit une loi normale $\mathcal{N}(m, \sigma^2)$. Intuitivement, on peut

calculer la moyenne empirique de l'échantillon et décider γ_0 si cette quantité est négative ou nulle, ou γ_1 dans le cas contraire. Puisque l'échantillon est simple, sa moyenne empirique $\overline{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ suit une loi normale $\mathcal{N}(m, \frac{\sigma^2}{n})$. La probabilité de l'erreur de première espèce est alors donnée par

$$\begin{aligned}
\alpha &= \text{proba}(\overline{X}_n > 0) \qquad \text{avec} \quad m \leq 0 \\
&= \int_0^{+\infty} \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{n(u-m)^2}{2\sigma^2}\Big)\, du \\
&= \int_{\frac{-m\sqrt{n}}{\sigma}}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{v^2}{2}\Big)\, dv
\end{aligned}$$

Cette valeur dépendant du paramètre inconnu m, il convient de considérer son maximum, obtenu pour $m = 0$, dans l'hypothèse où m est bel et bien négative ou nulle.

$$\alpha = \int_0^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{v^2}{2}\Big)\, dv = \frac{1}{2}$$

Cela signifie qu'une telle procédure nous conduira à rejeter à tort l'hypothèse H_0 dans la moitié des cas ! Pour réduire cette valeur excessive, la seule solution consiste à tolérer que la moyenne empirique excède 0 et atteigne une borne supérieure qui dépend totalement du seuil α. Par exemple, pour que ce seuil soit égal à 10%, il faut permettre à la moyenne empirique de prendre ses valeurs dans l'intervalle $]-\infty, a]$, où a est donné par

$$\begin{aligned}
0,1 &= \int_a^{+\infty} \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{n(u-m)^2}{2\sigma^2}\Big)\, du \\
&= \int_{\frac{(a-m)\sqrt{n}}{\sigma}}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{v^2}{2}\Big)\, dv \\
&\leq \int_{\frac{a\sqrt{n}}{\sigma}}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{v^2}{2}\Big)\, dv \qquad \text{pour} \quad m \in]-\infty, 0]
\end{aligned}$$

Les tables de la loi normale centrée réduite nous donnent le quantile à 90 %

$$\frac{a\sqrt{n}}{\sigma} \approx 1,285 \qquad \Rightarrow \qquad a \approx 1,285\ \frac{\sigma}{\sqrt{n}}$$

Enfin, il convient d'évaluer la puissance du test dans ces conditions (probabilité de rejeter à juste titre l'hypothèse H_0)

$$\begin{aligned}
\beta &= \int_a^{+\infty} \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{n(u-m)^2}{2\sigma^2}\Big)\, du \\
&= \int_{\frac{(a-m)\sqrt{n}}{\sigma}}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{v^2}{2}\Big)\, dv \\
&> \int_{\frac{a\sqrt{n}}{\sigma}}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{v^2}{2}\Big)\, dv \qquad \text{pour} \quad m \in]0, +\infty[\\
\beta &> 0,1
\end{aligned}$$

On constate évidemment dans ces conditions que la puissance du test est supérieure ou égale au seuil.

Cet exemple de test paramètrique, moyenne négative contre moyenne positive, fait partie des plus simples. Cependant le principe est toujours le même : il faut d'abord choisir une statistique dépendant du paramètre sur lequel porte le test. Ensuite, il convient de déterminer un intervalle dans lequel la probabilité que la statistique prenne ses valeurs, sous l'hypothèse H_0, soit égale au seuil α. Enfin, si c'est possible sous la contrainte précédente, il faut maximiser la puissance du test β.
La liste des tests paramétriques étant assez longue, consultez la littérature pour plus de détails sur ces tests d'hypothèses ! Ils font généralement appel aux lois de probabilité normale, de Student, du Khi-deux, de Fisher-Snedecor...

3.3.4 Stratégie Bayesienne

La seconde méthode minimise un **risque moyen**, à savoir une fonction qui prend en considération tout l'a priori dont on dispose. En voici un exemple.
Un émetteur numérique transmet une suite de symboles binaires 0 ou 1. En réception, arrivent des valeurs bruitées comprises, par exemple, entre -0,5 et 1,5. S'il parait logique de décider 1 lorsque l'on reçoit 0,8, quelle décision faut-il prendre quand on apprend par ailleurs que les 0 sont 30 fois plus fréquents que les 1 ? Pour répondre à une telle question, il faut intégrer tout l'a priori dont on dispose quant à la chaine de transmission. La modélisation globale s'écrit comme suit.
Soit $\underline{X} = \{X_1, \ldots, X_n\}$ un n-échantillon prélevé lors de la réception d'un seul symbole. On va tester deux hypothèses, H_0 et H_1, l'une contre l'autre. Par exemple, H_0 correspond à l'émission d'un 0, tandis que H_1 correspond à l'envoi d'un 1. On note p_0 et $p_1 = 1 - p_0$ les probabilités a priori de ces deux possibilités (elles ne sont pas nécessairement égales à 0,5 !). L'objectif du test d'hypothèses consiste en la mise en place d'une frontière pour partitionner l'ensemble des réalisations de l'échantillon, $\mathbb{R}^n$, en deux sous-ensembles G_0 et G_1 tels que, si $\underline{X}$ tombe dans G_0 on accepte H_0 (décision γ_0). Bien-sûr, si $\underline{X}$ tombe dans G_1, on accepte H_1 (décision γ_1).
Quatre cas de figure se présentent, auxquels on associe un coût :
- H_0 est vraie et on décide γ_0 : coût π_{00}
- H_0 est vraie et on décide γ_1 : coût π_{01} (erreur de $1^{ère}$ espèce)
- H_1 est vraie et on décide γ_0 : coût π_{10} (erreur de 2^{nde} espèce)
- H_1 est vraie et on décide γ_1 : coût π_{11}.

décision réalité	γ_0	γ_1
H_0	π_{00}	π_{01}
H_1	π_{10}	π_{11}

Ces quatre coûts dépendent bien-sûr du contexte, mais il est de bon ton de ne pas se tromper. On peut ainsi considérer que les coûts associés à une même hypothèse sont plus faibles quand la décision est bonne que quand elle est mauvaise :

$$\begin{cases} \pi_{00} & \leq & \pi_{01} \\ \pi_{11} & < & \pi_{10} \end{cases}$$

Enfin, on définit le "risque moyen" R que l'on va s'efforcer de minimiser :

$$R = \sum_{i,k} \pi_{ik}\ p(H_i, \gamma_k)\ \geq\ 0 \tag{3.19}$$

Or, $p(H_i, \gamma_0) = p(H_i) - p(H_i, \gamma_1)$ puisque G_0 et G_1 forment une partition de $\mathbb{R}^n$. On peut donc développer le risque comme suit :

$$\begin{aligned} R &= \pi_{00}[p(H_0) - p(H_0, \gamma_1)] + \pi_{01} p(H_0, \gamma_1) + \pi_{10}[p(H_1) - p(H_1, \gamma_1)] + \pi_{11} p(H_1, \gamma_1) \\ &= p_0\{\pi_{00}[1 - p(\gamma_1 \mid H_0)] + \pi_{01} p(\gamma_1 \mid H_0)\} + p_1\{\pi_{10}[1 - p(\gamma_1 \mid H_1)] + \pi_{11} p(\gamma_1 \mid H_1)\} \\ &= p_0\pi_{00} + p_0\{\pi_{01} - \pi_{00}\}\ p(\gamma_1 \mid H_0)] + p_1\pi_{10} + p_1\{\pi_{11} - \pi_{10}\}\ p(\gamma_1 \mid H_1) \\ &= p_0\pi_{00} + p_1\pi_{10} + \{p_0\ [\pi_{01} - \pi_{00}] p(\gamma_1 \mid H_0)\ -\ p_1\ [\pi_{10} - \pi_{11}] p(\gamma_1 \mid H_1)\} \\ &= p_0\pi_{00} + p_1\pi_{10} + \int_{G_1} \Big\{p_0\ [\pi_{01} - \pi_{00}] f_{\underline{X}}(\underline{x} \mid H_0) - p_1\ [\pi_{10} - \pi_{11}] f_{\underline{X}}(\underline{x} \mid H_1)\Big\} d\underline{x} \\ &= p_0\pi_{00} + p_1\pi_{10} + \int_{G_1} g(\underline{x}) d\underline{x} \end{aligned} \tag{3.20}$$

si l'on définit l'intégrande g

$$g(\underline{x}) = p_0\ [\pi_{01} - \pi_{00}] f_{\underline{X}}(\underline{x} \mid H_0) - p_1\ [\pi_{10} - \pi_{11}] f_{\underline{X}}(\underline{x} \mid H_1) \tag{3.21}$$

Dans l'expression (3.20), seule l'intégrale dépend de la partition retenue. La minimisation du risque s'en déduit aisément puisqu'il suffit de placer dans G_1 (respectivement G_0) tous les éléments de $\mathbb{R}^n$ pour lesquels l'intégrande $g(\underline{x})$ est négative (resp. positive) :

$$G_1 = \{\underline{x} \quad \text{tels que} \quad p_0[\pi_{01} - \pi_{00}]\ f_{\underline{X}}(\underline{x} \mid H_0)\ -\ p_1\ [\pi_{10} - \pi_{11}]\ f_{\underline{X}}(\underline{x} \mid H_1)\ <\ 0\} \tag{3.22}$$

Ce qui peut se réécrire selon

$$\frac{p_1}{p_0}\ \frac{f_{\underline{X}}(\underline{x} \mid H_1)}{f_{\underline{X}}(\underline{x} \mid H_0)}\ \begin{matrix} \gamma_1 \\ > \\ < \\ \gamma_0 \end{matrix}\ \frac{\pi_{01} - \pi_{00}}{\pi_{10} - \pi_{11}} \tag{3.23}$$

C'est la stratégie de Bayes, qui compare le "rapport de vraisemblance généralisée" à un seuil, fonction des coûts prédéfinis π_{ik}.
Un calcul analogue permet d'établir le risque moyen en fonction du sous-ensemble G_0 :

$$\begin{aligned} R &= p_1\pi_{11} + p_0\pi_{01} + \int_{G_0} \Big\{p_1\ [\pi_{10} - \pi_{11}] f_{\underline{X}}(\underline{x} \mid H_1) - p_0\ [\pi_{01} - \pi_{00}] f_{\underline{X}}(\underline{x} \mid H_0)\Big\} d\underline{x} \\ &= p_1\pi_{11} + p_0\pi_{01} - \int_{G_0} g(\underline{x}) d\underline{x} \end{aligned} \tag{3.24}$$

On constate naturellement que l'intégrande qui apparaît dans cette expression est l'opposée de celle qui figure dans l'équation (3.20). Ajoutons ces deux résultats, en tenant compte du fait que G_0 et G_1 forment une partition de $\mathbb{R}^n$, définie selon le signe de $g(\underline{x})$

$$\begin{aligned} 2R &= p_1\pi_{11} + p_0\pi_{01} + p_0\pi_{00} + p_1\pi_{10} - \int_{G_0} g(\underline{x}) d\underline{x} + \int_{G_1} g(\underline{x}) d\underline{x} \\ &= p_1(\pi_{10} + \pi_{11}) + p_0(\pi_{01} + \pi_{00}) - \int_{\mathbb{R}^n} |g(\underline{x})|\ d\underline{x} \end{aligned} \tag{3.25}$$

Ce résultat établit clairement que, plus l'intégrande g est contrastée, plus le risque moyen diminue. En particulier, si par malheur la fonction g était nulle partout, le risque moyen serait maximum (la prise de décision ne dépendant pas des observations recueillies).
Pour terminer, il est également intéressant d'établir une dernière formulation du risque (3.20) selon les probabilités de **fausse alarme** (notée F) et de **détection** (notée D)

$$\begin{cases} F &= \int_{G_1}\ f_{\underline{X}}(\underline{x} \mid H_0)\ d\underline{x} \\ D &= \int_{G_1}\ f_{\underline{X}}(\underline{x} \mid H_1)\ d\underline{x} \end{cases} \tag{3.26}$$

Alors

$$R = p_0\pi_{00} + (1 - p_0)\pi_{10} + p_0\ [\pi_{01} - \pi_{00}]F - (1 - p_0)\ [\pi_{10} - \pi_{11}]D \tag{3.27}$$

On constate que le risque augmente avec les fausses alarmes, ou lorsque la détection diminue.
Deux cas particuliers méritent d'être signalés :

– "Maximum de probabilité A Posteriori" :
Lorsque la fraction relative aux coûts vaut 1 ($\frac{\pi_{01}-\pi_{00}}{\pi_{10}-\pi_{11}} = 1$), la décision prise est conforme à la probabilité a posteriori maximale. En effet

$$p(H_k \mid x) = \frac{p(H_k, x)}{f_X(x)} = \frac{f_X(x \mid H_k)\ p(H_k)}{f_X(x)}$$

On choisit γ_1 si $p(H_1 \mid x) \ \geq \ p(H_0 \mid x)$.
– "Maximum de Vraisemblance" :
Lorsque la fraction relative aux coûts vaut 1 et les probabilités a priori sont égales ($p_1 = p_0 = \frac{1}{2}$), on décide selon le "rapport de vraisemblance" que l'on compare à 1.

3.3.5 Stratégie de Neymann et Pearson

Lorsque les probabilités a priori p_0 et p_1 sont inconnues, on a volontiers recours à la stratégie de Neymann et Pearson, où l'on recherche G_1 qui maximise la puissance β sous la contrainte $\alpha = s$, s étant une valeur choisie dans l'intervalle $[0, 1]$. Là encore, la décision découle de la comparaison du rapport de vraisemblance à un seuil, fonction du paramètre s.
Soient T un sous-ensemble de $\mathbb{R}^n$ tel que $\int_T f_{\underline{X}}(\underline{x} \mid H_0)\ d\underline{x} \ \leq \ s$
et $G_1 = \left\{ \underline{x} \in \mathbb{R}^n \ / \ f_{\underline{X}}(\underline{x} \mid H_1) \geq \lambda f_{\underline{X}}(\underline{x} \mid H_0) \right\}$
sous la contrainte $\int_{G_1} f_{\underline{X}}(\underline{x} \mid H_0)\ d\underline{x} = s$
On note C l'intersection, éventuellement vide, de T et G_1, puis B et A leurs compléments respectifs

$$\begin{cases} G_1 &= A \cup C \quad , \quad A \cap C &= \emptyset \\ T &= B \cup C \quad , \quad B \cap C &= \emptyset \end{cases}$$

$$\begin{aligned}
\int_{G_1} f_{\underline{X}}(\underline{x} \mid H_1)\ d\underline{x} &= \int_A f_{\underline{X}}(\underline{x} \mid H_1)\ d\underline{x} + \int_C f_{\underline{X}}(\underline{x} \mid H_1)\ d\underline{x} \\
&\geq \int_A \lambda f_{\underline{X}}(\underline{x} \mid H_0)\ d\underline{x} + \int_C f_{\underline{X}}(\underline{x} \mid H_1)\ d\underline{x} \\
&= \int_{G_1} \lambda\ f_{\underline{X}}(\underline{x} \mid H_0)\ d\underline{x} - \int_C \lambda\ f_{\underline{X}}(\underline{x} \mid H_0)\ d\underline{x} + \int_C f_{\underline{X}}(\underline{x} \mid H_1)\ d\underline{x} \\
&\geq \int_T \lambda\ f_{\underline{X}}(\underline{x} \mid H_0)\ d\underline{x} - \int_C \lambda\ f_{\underline{X}}(\underline{x} \mid H_0)\ d\underline{x} + \int_C f_{\underline{X}}(\underline{x} \mid H_1)\ d\underline{x} \\
&= \int_B \lambda f_{\underline{X}}(\underline{x} \mid H_0)\ d\underline{x} + \int_C f_{\underline{X}}(\underline{x} \mid H_1)\ d\underline{x} \\
&\geq \int_B f_{\underline{X}}(\underline{x} \mid H_1)\ d\underline{x} + \int_C f_{\underline{X}}(\underline{x} \mid H_1)\ d\underline{x} \\
&= \int_T f_{\underline{X}}(\underline{x} \mid H_1)\ d\underline{x} \qquad (3.28)
\end{aligned}$$

Il en résulte que, quel que soit T de même niveau de signification du test ($\alpha = s$) que G_1, sa puissance est inférieure ou égale à celle établie sur G_1. En d'autres termes, à seuil α donné, G_1 définit le test de puissance maximale.
On vérifie alors que, quel que soit le réel positif λ, la puissance β est nécessairement supérieure au niveau de signification fixé a priori $\alpha = s$. En effet

$$\begin{cases} \beta &= \int_{G_1(\lambda)} f_{\underline{X}}(\underline{x} \mid H_1)\ d\underline{x} = 1 - \int_{G_0(\lambda)} f_{\underline{X}}(\underline{x} \mid H_1)\ d\underline{x} \\ \alpha &= \int_{G_1(\lambda)} f_{\underline{X}}(\underline{x} \mid H_0)\ d\underline{x} = 1 - \int_{G_0(\lambda)} f_{\underline{X}}(\underline{x} \mid H_0)\ d\underline{x} \end{cases} \qquad (3.29)$$

avec la partition de $\mathbb{R}^n$ selon les deux sous-ensembles

$$\begin{cases} G_1(\lambda) &= \{\ \underline{x} \in \mathbb{R}^n \ / \ f_{\underline{X}}(\underline{x} \mid H_1) \ \geq \ \lambda f_{\underline{X}}(\underline{x} \mid H_0)\ \} \\ G_0(\lambda) &= \{\ \underline{x} \in \mathbb{R}^n \ / \ f_{\underline{X}}(\underline{x} \mid H_1) \ < \ \lambda f_{\underline{X}}(\underline{x} \mid H_0)\ \} \end{cases}$$

Deux cas se présentent, selon la valeur du paramètre λ
– $\lambda \geq 1$: $\beta = \int_{G_1(\lambda)} f_{\underline{X}}(\underline{x} \mid H_1)\ d\underline{x} \ \geq \ \int_{G_1(\lambda)} \lambda f_{\underline{X}}(\underline{x} \mid H_0)\ d\underline{x} = \lambda\alpha \ \geq \ \alpha$
– $\lambda \leq 1$: $1 - \beta = \int_{G_0(\lambda)} f_{\underline{X}}(\underline{x} \mid H_1)\ d\underline{x} \ < \ \int_{G_0(\lambda)} \lambda f_{\underline{X}}(\underline{x} \mid H_0)\ d\underline{x} = \lambda(1-\alpha) \ \leq \ 1 - \alpha$
Dans les deux cas, on a bien $\beta \geq \alpha$.

Bibliographie

[1] BASS, “Eléments de calcul des probabilités”, Masson

[2] BOROVKOV, “Statistique mathématique”, Editions MIR

[3] JAFFARD, “Statistique”, Masson

[4] KOROLIOUK, “Aide mémoire de théorie des probabilités et de statistique mathématique”, Editions MIR

[5] LEVINE, “Radiotechnique statistique (tome 1)”, Editions MIR

[6] LEVINE, “Radiotechnique statistique (tome 2)”, Editions MIR

[7] RENYI, “Calcul des probabilités”, Jacques Gabay

[8] VENTSEL, “Théorie des probabilités”, Editions MIR

[9] SPIEGEL, “Probabilités et statistiques”, Série Schaum