



Eléments de statistique inférentielle

par :

Faïcel CHAMROUKHI

`chamroukhi@univ-tln.fr`

`http://chamroukhi.univ-tln.fr`

Table des matières

1	Introduction et principes généraux	5
1.1	Introduction	5
1.1.1	Notion de population, échantillon, individu	5
1.1.2	Expérience aléatoire	6
1.2	Variables aléatoires	6
1.2.1	Types de variables aléatoires	6
1.2.2	Fonction de densité de probabilité	7
1.2.3	Fonction de répartition	7
1.2.4	Espérance mathématique et variance d'une variable aléatoire	7
1.2.5	Espérance mathématique d'une variable aléatoire discrète	8
1.2.6	Espérance mathématique d'une variable aléatoire continue	8
1.2.7	Variance d'une variable aléatoire	8
2	Couple de variables aléatoires, notion d'indépendance et vecteurs aléatoires	11
2.1	Couple de variables aléatoires et notion d'indépendance	11
2.1.1	Lois jointes, lois conditionnelles et théorème de Bayes	11
2.1.2	Loi et densité de couple de variables aléatoires	11
2.1.3	Indépendance de deux variables aléatoires	13
2.1.4	Indépendance et espérance	14
2.1.5	Loi des grands nombres	15
3	Vecteurs Aléatoires	17
3.1	Espérance d'un vecteur aléatoire	18
3.2	Matrice de covariance	18
3.3	Matrice de corrélation	19
3.4	Indépendance de vecteurs aléatoires	19
3.4.1	Cas des variables discrètes	19
3.4.2	Cas des variables aléatoires à densité	19
3.5	Quelques statistique sur vecteurs aléatoires	19
4	Variables aléatoires gaussiennes et vecteurs aléatoires gaussiens	21
4.1	Loi normale mono-dimensionnelle	21
4.1.1	Loi normale générale	21
4.1.2	Loi normale centrée réduite	21
4.1.3	Fonction de répartition	22
4.1.4	Quelques propriétés	22
4.2	Théorème central limite	23
4.3	Couple de variables aléatoires gaussiennes	24

4.4	Vecteurs aléatoires gaussiens	24
4.4.1	Loi normale multidimensionnelle	24
4.4.2	Quelques propriétés	25
5	Estimation de paramètres	27
5.1	Introduction	27
5.2	Critères de qualité pour les estimations	27
5.2.1	Cas de plusieurs paramètres	29
5.3	Méthodes d'estimation	31
6	Méthode du maximum de vraisemblance	33
6.1	Définition de la fonction de vraisemblance	33
6.2	Maximum de vraisemblance	34
6.2.1	Cas d'un seul paramètre à estimer	34
6.2.2	Cas de plusieurs paramètres à estimer (Vraisemblance multi-variée)	35
6.3	Propriétés	35
6.4	Cas gaussien	36
6.5	Estimation par intervalle	37
6.6	Intervalle de confiance	37
6.7	Cas d'estimation d'une gaussienne	37
6.7.1	Intervalle de confiance pour μ dans $\mathcal{N}(\mu, \sigma^2)$ avec σ connu	37
7	Méthode des Moindres Carrés	39
7.1	Méthode des Moindres Carrés	39
7.1.1	Définition des Moindres Carrés	39
7.1.2	Moindres Carrés	40
7.1.3	Propriétés de l'estimateur des moindres carrés	41
8	Régression linéaire	43
8.1	Introduction	43
8.2	Le modèle linéaire simple	43
8.2.1	Estimation par moindres carrés	44
8.2.2	Formulation vectorielle	46
9	Estimation par intervalle	49
9.1	Intervalle de confiance	49
9.1.1	Calcul d'un intervalle de confiance	50
9.1.2	Loi normale : Intervalle de confiance sur μ	50
9.1.3	Loi normale : Intervalle de confiance sur σ^2	52
10	Tests d'hypothèses	53
10.1	Région de rejet d'un test	53
10.2	Erreurs associées à un test	54
10.3	Statistiques de test	55
10.3.1	Test du rapport de vraisemblance	55
10.3.2	Test de Wald	56

Chapitre 1

Introduction et principes généraux

1.1 Introduction

Dans ce chapitre, nous présenterons/rappellerons des principes et concepts généraux en probabilités et statistiques, des outils nécessaires pour comprendre le fondement de l'inférence statistique.

1.1.1 Notion de population, échantillon, individu

Une **population statistique** (qu'on note Ω) est un ensemble concernée par une étude statistique. Un **individu** est un élément individuel considéré dans l'analyse statistique. Un **échantillon** est un sous-ensemble effectivement observé de la population. La taille n de l'échantillon est le cardinal de ce sous-ensemble considérée dans l'étude statistique.

⚠ Remarque 1.1.1. *Dans la communauté du traitement de signal, on appelle aussi échantillon un point d'un signal, ce qui correspond donc, dans certains cas, à un individu en analyse statistique.*

Il existe principalement deux types d'approches statistiques : **descriptive** et **inférentielle**.

Statistique descriptive :

Définition 1.1.1. *L'objectif de la statistique descriptive est de décrire les données disponibles dans un **échantillon** par descriptions les plus synthétiques possibles.*

Ceci peut être effectué par des représentations graphiques, , résumés numériques, etc. Les loi de probabilités ne jouent pas un grand rôle dans cette approche.

Statistique inférentielle :

Définition 1.1.2. *L'inférence statistique consiste à induire (inférer) les caractéristiques inconnues d'une **population** à partir d'un échantillon issu de cette population*

Le modèles probabilistes sont au cœur de cette approche.

1.1.2 Expérience aléatoire

On appelle **expérience aléatoire** une expérience qui, répétée plusieurs fois dans des conditions opératoires identiques, produit des résultats qui peuvent être différents. L'ensemble de tous les résultats possibles est appelé *univers*. Un événement aléatoire correspond à une affirmation qui peut être vraie ou fausse suivant le résultat de l'expérience aléatoire. Un *événement aléatoire* est donc représenté par l'ensemble de résultats de l'expérience aléatoire (un sous-ensemble de l'univers).

La notion d'expérience aléatoire \mathcal{E} se formalise mathématiquement en définissant :

1. l'ensemble fondamental Ω (appelé l'*univers*) définissant l'ensemble des résultats possibles de \mathcal{E} , appelés événements élémentaires ;
2. un ensemble \mathcal{A} de parties de Ω , appelées événements.
3. une fonction $\mathbb{P} : \mathcal{A} \rightarrow [0, 1]$, appelée mesure ou distribution de probabilité, qui à tout événement A associe un nombre $\mathbb{P}(A)$ appelé probabilité de cet événement.

1.2 Variables aléatoires

Une **Variable aléatoire** (statistique) X est une application de Ω dans un ensemble E

$$X : \Omega \longrightarrow E$$

On en trouve plusieurs types de variables aléatoires.

1.2.1 Types de variables aléatoires

Soit X un variable aléatoire. X est dite une **variable quantitative** si elle prend des valeurs dans \mathbb{R} . Il s'agit d'une variable représentée par une quantité (une valeur) telle que l'âge, le poids et la taille, etc.

Une **variable qualitative** est quant à elle une variable représentée par une qualité (ou catégorie), telles que le sexe, encore l'état civil, le degré de satisfaction d'un service quelconque, etc. La variable qualitative est donc représenté par une modalité plutôt que d'une valeur.

Il existe deux types de variables de variables quantitatives : **variables discrètes** et **variables continues**.

Définition 1.2.1. Une variable aléatoire est dite **discrète** si elle ne prend que des valeurs discontinues dans un ensemble fini ou dénombrable.

Exemples : le résultat du jet d'un dé, le nombre d'enfants dans une famille, le nombre de personnes ayant une maladie rare dans une ville donnée, sont des variables aléatoires discrètes.

Définition 1.2.2. Une variable aléatoire est dite **continue** ou **a densité** si elle ne prend ses valeurs dans \mathbb{R} ou une partie ou un ensemble de parties de \mathbb{R} .

Exemple : la moyenne des étudiants, la taille, etc sont des variables aléatoires continues.

Pour les variable aléatoires qualitatives, il existe également deux types de variables : **variables nominales** et **variables ordinales**.

Définition 1.2.3. Une variable aléatoire qualitative **nominale** est une variable qui correspond à des noms, il n'y a aucun ordre précis sur les modalités. Ce sont seulement des mots dans le désordre.

Par exemple, la variable sexe est une variable qualitative nominale qui a deux modalités possibles : féminin ou masculin et dont l'ordre n'importe pas. Un autre exemple est celui de la catégorie socioprofessionnelle

1.2.2 Fonction de densité de probabilité

On appelle densité de probabilité toute fonction continue par morceaux (intégrable) :

$$\begin{aligned} f : \mathbb{R} &\rightarrow \mathbb{R}^+ \\ x &\rightarrow f(x) \end{aligned} \quad (1.1)$$

telle que :

$$\begin{aligned} (1) \quad &\forall x \in \mathbb{R} \quad f(x) \geq 0 \\ (2) \quad &\int_{\mathbb{R}} f(x) dx = 1 \end{aligned}$$

1.2.3 Fonction de répartition

La fonction de répartition d'une variable aléatoire réelle X , notée F_X , caractérise la loi de probabilité de cette variable. La fonction de répartition F_X associée à tout réel x

$$F_X(x) = \mathbb{P}(X \leq x) \quad (1.2)$$

qui représente la probabilité que la variable aléatoire réelle X prenne une valeur inférieure ou égale à x .

La fonction de répartition F_X d'une variable aléatoire X de densité de probabilité f_X est définie, pour tout nombre réel x , par :

$$F_X(x) = \int_{-\infty}^x f_X(u) du. \quad (1.3)$$

La fonction de répartition $F_X(x)$ est donc l'une des primitives de la fonction de densité de probabilités.

⚠ Remarque 1.2.1. La probabilité que X se trouve dans l'intervalle $]a, b]$ est donc, si $a < b$,

$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$$

1.2.4 Espérance mathématique et variance d'une variable aléatoire

L'espérance mathématique d'une variable aléatoire réelle représente la valeur moyenne prise par cette variable aléatoire.

1.2.5 Espérance mathématique d'une variable aléatoire discrète

Soit X une variable aléatoire discrète prenant ses valeurs dans l'ensemble $\{x_1, \dots, x_n\}$ avec des probabilités respectives p_1, \dots, p_n c.à.d $P(X = x_k) = p_k$ et $\sum_{k=1}^n p_k = 1$. L'espérance de X est donnée par

$$\mathbb{E}[X] = \sum_k x_k P(X = x_k) = \sum_k x_k p_k \quad (1.4)$$

△ Remarque 1.2.2. On peut remarquer que l'espérance d'une variable aléatoire binaire est égale à sa probabilité de valoir 1, autrement dit si X est une v.a binaire (prenant ses valeurs dans $\{0, 1\}$, on a alors $\mathbb{E}[X] = P(X = 1)$).

1.2.6 Espérance mathématique d'une variable aléatoire continue

Soit X une variable aléatoire continue à valeurs réelles ayant comme fonction de densité de probabilité f_X . L'espérance de X est donnée par

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f_X(x) dx. \quad (1.5)$$

Plus généralement, soit $g : \mathbb{R} \rightarrow \mathbb{R}$, l'espérance de $g(X)$ est donnée par

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x) f_X(x) dx \quad (1.6)$$

1.2.7 Variance d'une variable aléatoire

On appelle variance d'une variable aléatoire X la quantité définie par :

$$\text{var}[X] = \mathbb{E}[(X - E[X])^2] = \mathbb{E}[X^2] - (E[X])^2 \quad (1.7)$$

Ainsi, la variance notée σ_X^2 et sa racine carrée l'écart-type (noté σ_X) mesurent la dispersion de la variable aléatoire X autour de son espérance $E[X]$.

△ Remarque 1.2.3. D'après 1.6, la variance est donc aussi donnée par

$$\text{var}[X] = \int_{\mathbb{R}} (x - E[X])^2 f_X(x) dx. \quad (1.8)$$

Quelques propriétés

Soit X une v.a réelle, deux fonctions $g, h : \mathbb{R} \rightarrow \mathbb{R}$ et deux réels a et b , alors

$$\mathbb{E}[a.g(X) + b.h(X)] = a.\mathbb{E}[g(X)] + b.\mathbb{E}[h(X)]$$

ainsi le cas particulier affine consiste en la propriété suivante :

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

⇒ On dit que l'espérance est **linéaire**

Soient n variables aléatoires réelles X_1, \dots, X_n . Alors $\mathbb{E}[\sum_i X_i] = \sum_i \mathbb{E}[X_i]$
De même, pour la variance, on a la propriété suivante :

$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

⇒ On dit que la variance a une propriété quadratique **quadratique**

⚠ Remarque 1.2.4. Plus généralement on parle de moment d'ordre r ($r > 0$) de la v.a X la quantité donnée par

$$\mathbb{E}[X^r] = \int_{\mathbb{R}} x^r f_X(x) dx$$

et respectivement de moment centré d'ordre r ($r > 0$) la quantité donnée par

$$\mathbb{E}[(X - \mathbb{E}[X])^r] = \int_{\mathbb{R}} (x - \mathbb{E}(X))^r f_X(x) dx.$$

Quelques statistique sur variables aléatoires

Soit (X_1, \dots, X_n) un échantillon donné d'une population ayant comme fonction de densité de probabilité la fonction f_X de paramètre θ (une distribution de probabilités $P_X(\cdot; \theta)$ pour le cas discret). Une *statistique* est toute fonction de l'échantillon donné (X_1, \dots, X_n) qui ne dépend pas de θ .

On en distingue deux statistiques classiquement utilisées et particulièrement intéressantes pour une étude statistique et qui sont : la *moyenne empirique*

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (1.9)$$

et la *variance empirique*

$$S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2. \quad (1.10)$$

La moyenne est la variance sont dit le moment d'ordre 1 et le moment d'ordre 2, respectivement. Plus généralement, on parle du *moment empirique d'ordre k* donné par :

$$M_k = \frac{1}{n} \sum_{i=1}^n X_i^k. \quad (1.11)$$

Chapitre 2

Couple de variables aléatoires, notion d'indépendance et vecteurs aléatoires

2.1 Couple de variables aléatoires et notion d'indépendance

2.1.1 Lois jointes, lois conditionnelles et théorème de Bayes

Soient X et Y deux variables aléatoires de densités respectives $f_X(x)$ et $f_Y(y)$ (ou de lois respectives $P(X)$ et $P(Y)$ pour le cas discret). Le comportement du couple (X, Y) est entièrement décrit par leur fonction de **densité de probabilité jointe** $f_{XY}(x, y)$ (ou **distribution jointe** de probabilité dans le cas discret $P_{XY}(X = x, Y = y)$). Il se peut que lors du tirage d'une réalisation de (X, Y) , que la valeur observée (réalisation) x de X fournisse une information sur la valeur possible de Y . Cette information est représentée par la distribution de Y conditionnellement à $X = x$ (distribution de Y étant donné à $X = x$) soit $f_{Y|X}(y|x)$. Ceci est explicité par le théorème de suivant, appelé théorème de Bayes (ou aussi règle de Bayes ou formule de Bayes) comme suit

Théorème 2.1.1.

$$f_{XY}(x, y) = f_{Y|X}(y|x)f_X(x) \quad (2.1)$$

et par symétrie on a également

$$f_{XY}(x, y) = f_{X|Y}(x|y)f_Y(y) . \quad (2.2)$$

Pour les VA discrètes cela revient à

$$P(Y = y, X = x) = P(Y = y|X = x)P(X = x) \quad (2.3)$$

$$= P(X = x|Y = y)P(Y = y) \quad (2.4)$$

2.1.2 Loi et densité de couple de variables aléatoires

Cas de v.a discrètes

Soient X et Y deux variables discrètes telle que X prend ses valeurs dans l'ensemble A et Y prend ses valeurs dans l'ensemble B . La loi du couple (X, Y) est

définie par l'ensemble des probabilités :

$$P(X = x, Y = y) \quad \text{avec } x \in A \text{ et } y \in B.$$

⚠ Remarque 2.1.1. Cette loi peut être représenté sous la forme d'un tableau dans le cas où les v.a prennent un nombre petit de valeurs

Y/X	...	x	...	Somme des colonnes
⋮				
y		$P(X = x, Y = y)$		$P(Y = y)$
⋮				
Somme des lignes		$P(X = x)$		

TABLE 2.1 – Loi jointe de couple de v.a discrètes

La loi de chaque variable est donc donnée par la loi marginale comme suit :

$$P(X = x) = \sum_{y \in B} P(X = x, Y = y) \quad \text{et} \quad P(Y = y) = \sum_{x \in A} P(X = x, Y = y) \quad (2.5)$$

On peut remarquer que ces lois correspondent respectivement à la somme des lignes et la somme des colonnes dans le tableau précédent (2.1).

Cas de v.a continues (à densité)

Un couple de v.a. réelles (X, Y) est dit à densité s'il existe une fonction $f_{(X,Y)}$ telle que la fonction de répartition du couple s'écrit

$$F_{X,Y}(x, y) = \int_{-\infty}^x \int_{-\infty}^y f_{X,Y}(u, v) du dv \quad (2.6)$$

satisfaisant les conditions suivantes :

1. $f_{X,Y}(x, y) \geq 0$ pour tout $(x, y) \in \mathbb{R}^2$,
2. $\int_{\mathbb{R}} \int_{\mathbb{R}} f_{X,Y}(x, y) dx dy = 1$.

Densité marginales

Les variables X et Y sont des variables continues de densité respectives

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x, y) dy \quad (2.7)$$

et

$$f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x, y) dx. \quad (2.8)$$

Covariance de deux variables aléatoires

Quand il s'agit d'une variable aléatoire, on parle de variance comme dans les définitions (1.7)(1.8). Quand il s'agit d'un couple de deux variables aléatoires différentes, on parle de **covariance** de deux variables aléatoires.

On définit la covariance de deux variables aléatoires X et Y comme le terme

$$\text{cov}(X, Y) \equiv \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \quad (2.9)$$

et on la propriété suivante

$$\text{cov}(X, Y) = \sigma_{XY} = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \quad (2.10)$$

L'espérance du produit $E(XY)$ se calcule à partir de la loi jointe du couple (X, Y) . Ainsi, dans le cas discret, si X prend ses valeurs dans A et Y prends ses valeurs dans B , on a

$$\mathbb{E}[XY] = \sum_{x \in A} \sum_{y \in B} xyP(X = x, Y = y) \quad (2.11)$$

Dans le cas de v.a réelles continu, on a

$$\mathbb{E}[XY] = \int_{\mathbb{R}} \int_{\mathbb{R}} xyf_{X,Y}(x, y) dx dy. \quad (2.12)$$

La covariance mesure la dépendance entre deux v.a. Si les deux v.a sont indépendantes, leur covariance est donc nulle.

Corrélation entre deux variables aléatoires

La corrélation entre deux variables aléatoires se mesure par le coefficient de corrélation linéaire noté ρ .

Définition 2.1.1. *Le coefficient de corrélation linéaire entre deux variables aléatoires X et Y est défini par :*

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}. \quad (2.13)$$

On a toujours $\rho_{X,Y} \in [-1, 1]$. Plus $|\rho_{X,Y}|$ est proche de 1 plus la dépendance entre les variables X et Y est forte.

▲ Remarque 2.1.2. *Si X et Y sont indépendantes, alors $\text{Cov}(X, Y) = 0$ et donc $\rho_{X,Y} = 0$. On a par conséquent, $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$.*

2.1.3 Indépendance de deux variables aléatoires

Soient X et Y deux variables aléatoires de densités respectives $f_X(x)$ et $f_Y(y)$ (ou de lois respectives $P(X)$ et $P(Y)$ pour le cas discret).

Définition 2.1.2. *X et Y sont dites indépendantes si et seulement si leur densité de probabilité jointe est égale au produit de leurs densités marginales. Plus spécifiquement*

$$f_{XY}(x, y) = f_X(x)f_Y(y) \quad (2.14)$$

Pour les VA discrètes cela revient à

$$P(Y = y, X = x) = P(X = x)P(Y = y) \quad (2.15)$$

Cela vient du fait que, d'une part le comportement du couple (X, Y) est entièrement décrit par leur fonction de densité de probabilité jointe (ou distribution jointe de probabilité dans le cas discret) $f_{XY}(x, y)$ et d'autre part, après le tirage d'une réalisation de (X, Y) , la valeur observée (réalisation) x de X fournit une certaine quantité d'information sur la valeur de Y . Cette information est représentée par la distribution de Y conditionnellement à $X = x$ soit $f_{Y|X}(y|x)$. X et Y sont dites indépendantes si observer une réalisation x de X n'a aucun effet et n'apporte aucune information sur la réalisation possible de Y étant donné x . Autrement dit, la distribution de Y conditionnellement à X ne dépend pas de x .

La densité de probabilité jointe $f_{XY}(x, y)$ est :

$$f_{XY}(x, y) = f_{Y|X}(y|x)f_X(x) \quad (2.16)$$

par le théorème de Bayes. Ensuite la densité marginale de Y est obtenue en intégrant cette densité jointe par rapport à x :

$$f_Y(y) = \int_{\mathcal{X}} f_{XY}(x, y) dx = \int_{\mathcal{X}} f_{Y|X}(y|x)f_X(x) \quad (2.17)$$

Mais si la densité conditionnelle de Y ne dépend pas de x (cas de VA indépendantes), on peut la sortir de l'intégrale et nous avons alors

$$f_Y(y) = f_{Y|X}(y|x) \int_{\mathcal{X}} f_X(x) = f_{Y|X}(y|x) \quad (2.18)$$

donc

$$f_{Y|X}(y|x) = f_Y(y) \quad (2.19)$$

et par conséquent

$$f_{XY}(x, y) = f_X(x)f_Y(y) \quad (2.20)$$

△ Remarque 2.1.3. Pour le cas des VA discrètes, le raisonnement est le même en remplaçant les intégrales par des sommes.

2.1.4 Indépendance et espérance

X et Y sont indépendantes si et seulement si pour toutes fonctions $f(X)$ et $g(Y)$

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$$

si les espérances existent. Un cas particulier est bien celui où $f(X) = X$ et $g(Y) = Y$. On peut alors énoncer le théorème ainsi Si X et Y sont indépendantes, alors

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

Une conséquence de ce théorème est que deux variables indépendantes sont décorrélées (leur covariance est nulle).

△ Remarque 2.1.4. La réciproque est fautive : la décorrélation de deux variables n'implique pas indépendance, sauf dans le cas v.a à densité normale comme on le verra plus tard dans le chapitre dédié aux variables aléatoires gaussiennes.

2.1.5 Loi des grands nombres

Convergence en probabilité : Définition On dit qu'une suite (X_n) de v.a. converge en probabilité vers une v.a. X si

$$\forall \epsilon > 0, \quad \lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0. \quad (2.21)$$

Théorème 2.1.2 (Loi des grands nombres). Soit X_1, \dots, X_n un échantillon indépendant d'une suite de v.a. indépendantes et de même loi d'espérance $\mathbb{E}[X] = \mu$ et de variance $\text{var}(X) = \sigma^2$. Alors quel on a :

$$\forall \epsilon > 0, \quad \lim_{n \rightarrow +\infty} \mathbb{P}\left(\left|\frac{X_1 + X_2 + \dots + X_n}{n} - \mathbb{E}(X)\right| \geq \epsilon\right) = 0 \quad (2.22)$$

\Rightarrow La moyenne empirique $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ converge en probabilité vers l'espérance $\mathbb{E}[X]$ quand $n \rightarrow +\infty$ (asymptotiquement)

⚠ Remarque 2.1.5. Ici on parle donc de *convergence en probabilité*.

\Rightarrow La L.G.N nous dit que, pour tout réel ϵ strictement positif, la probabilité que la moyenne empirique $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ s'éloigne de l'espérance $\mathbb{E}[X] = \mu$ d'au moins ϵ tend vers 0 quand la taille de l'échantillon n tend vers l'infini.

Chapitre 3

Vecteurs Aléatoires

Un vecteur aléatoire réel \mathbf{X} de dimension d est un vecteur $\mathbf{X} = (X_1, \dots, X_d)^T$ dont les composantes $X_j, j = 1, \dots, d$ sont des variables aléatoires réelles.

La fonction de répartition d'un vecteur aléatoire \mathbf{X} décrit la loi de probabilité d'un vecteur aléatoire et est donnée par

$$F_{\mathbf{X}}(\mathbf{x}) = F_{\mathbf{X}}(x_1, \dots, x_d) = P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_d \leq x_d) \quad (3.1)$$

Dans le cas de vecteurs aléatoires réels continus (ayant des composantes continues), la fonction de densité de probabilité $f : \mathbb{R}^d \rightarrow [0, +\infty[$ de \mathbf{X} est définie par

$$f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}}(x_1, \dots, x_d) = \frac{\partial^d F_{\mathbf{X}}(\mathbf{x})}{\partial x_1 \partial x_2 \dots \partial x_d}. \quad (3.2)$$

Si cette f.d.p existe, on alors

$$\mathbb{P}(\mathbf{X} \in A) = \int_A f_{\mathbf{X}}(\mathbf{x}) \, d\mathbf{x}, \quad (3.3)$$

pour tout $A \subseteq \mathbb{R}^d$ pour lequel cette intégrale est définie, et comme $f_{\mathbf{X}}$ est une f.d.p, on a :

$$\int_{\mathbb{R}^d} f_{\mathbf{X}}(\mathbf{x}) \, d\mathbf{x} = 1. \quad (3.4)$$

Lorsque \mathbf{X} est un vecteur aléatoire discret, la loi de probabilité (fonction de masse de probabilité) se définit par $P_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(\mathbf{X} = \mathbf{x})$. On a aussi

$$\mathbb{P}(\mathbf{X} \in A) = \sum_{\mathbf{x} \in A} P_{\mathbf{X}}(\mathbf{x}), \quad (3.5)$$

et

$$\sum_{\mathbf{x} \in \mathcal{X}} P_{\mathbf{X}}(\mathbf{x}) = 1, \quad (3.6)$$

\mathcal{X} étant l'ensemble de valeurs que prend \mathbf{x} .

3.1 Espérance d'un vecteur aléatoire

Soit $\mathbf{X} = (X_1, \dots, X_d)^T$ un vecteur aléatoire réel. Notons par $\mu_j = \mathbb{E}[X_j]$ l'espérance de chaque v.a composante X_j .

Définition 3.1.1. L'espérance du vecteur aléatoire \mathbf{X} est donnée par le vecteur déterministe

$$\mathbb{E}[\mathbf{X}] = (\mathbb{E}[X_1], \dots, \mathbb{E}[X_d])^T = (\mu_1, \dots, \mu_d)^T$$

3.2 Matrice de covariance

La matrice de variance-covariance (appelée aussi matrice de covariance) d'un vecteur aléatoire \mathbf{X} est la matrice carrée parfois notée Σ dont le terme générique est donné par :

$$\Sigma_{i,j} = \text{cov}(X_i, X_j)$$

Elle est définie comme :

$$\Sigma_{\mathbf{X}} = \text{var}(\mathbf{X}) = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T], \quad (3.7)$$

et donc donnée aussi par

$$\Sigma_{\mathbf{X}} = \mathbb{E}[\mathbf{X}\mathbf{X}^T] - \mathbb{E}[\mathbf{X}](\mathbb{E}[\mathbf{X}])^T, \quad (3.8)$$

$\mathbb{E}[\mathbf{X}]$ étant l'espérance mathématique de \mathbf{X} . En développant les termes on obtient la forme suivante :

$$\begin{aligned} \Sigma &= \begin{bmatrix} \mathbb{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \mathbb{E}[(X_1 - \mu_1)(X_2 - \mu_2)] & \cdots & \mathbb{E}[(X_1 - \mu_1)(X_n - \mu_n)] \\ \mathbb{E}[(X_2 - \mu_2)(X_1 - \mu_1)] & \mathbb{E}[(X_2 - \mu_2)(X_2 - \mu_2)] & \cdots & \mathbb{E}[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[(X_n - \mu_n)(X_1 - \mu_1)] & \mathbb{E}[(X_n - \mu_n)(X_2 - \mu_2)] & \cdots & \mathbb{E}[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix} \\ &= \begin{pmatrix} \text{var}(X_1) & \text{cov}(X_1 X_2) & \cdots & \text{cov}(X_1 X_p) \\ \text{cov}(X_2 X_1) & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(X_p X_1) & \cdots & \cdots & \text{var}(X_p) \end{pmatrix} = \begin{pmatrix} \sigma_{X_1}^2 & \sigma_{X_1 X_2} & \cdots & \sigma_{X_1 X_p} \\ \sigma_{X_2 X_1} & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_p X_1} & \cdots & \cdots & \sigma_{X_p}^2 \end{pmatrix} \end{aligned}$$

où les $\sigma_{X_j}^2, j = 1, \dots, p$ sont les variances respectives des v.a X_j et les $\sigma_{X_i, X_j}^2, i \neq j$ sont les covariances respectives des couples de v.a $(X_i, X_j) : \sigma_{X_i, X_j}^2 = \text{cov}(X_i, X_j)$.

La matrice de covariance possède les propriétés suivantes :

1. La matrice est **symétrique** car on a $\text{cov}(X, Y) = \text{cov}(Y, X)$
2. Elle est **semi-définie positive** (ses valeurs propres sont positives ou nulles).
3. Les éléments de sa diagonale ($\Sigma_{i,i}$) représentent la variance de chaque variable, étant donné la propriété $\text{cov}(X, X) = \text{var}(X)$
4. Les éléments en dehors de la diagonale ($\Sigma_{i,j}, i \neq j$) représentent la covariance entre les variables i et j .
5. etc

3.3 Matrice de corrélation

La matrice de corrélation (notée R) du vecteur aléatoire \mathbf{X} , définie de manière analogue à la matrice de covariance, est la matrice dont le terme général est le coefficient de corrélation linéaire $\rho_{i,j}$ donné par l'équation (2.13). Chaque terme $\rho_{i,j}$ représente la corrélation entre le couple de variables (X_i, X_j) du vecteur \mathbf{X} et pour rappel est donné par $\rho_{i,j} = \frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{Var}(X_i)\text{Var}(X_j)}} = \frac{\sigma_{i,j}}{\sigma_i\sigma_j}$.

⚠ Remarque 3.3.1. Toutes les valeurs de la matrice de corrélation sont donc comprises entre -1 et $+1$ et les termes de la diagonale sont égaux à 1 .

3.4 Indépendance de vecteurs aléatoires

notion de VA i.i.d

3.4.1 Cas des variables discrètes

Soit $\mathbf{X} = (X_1, X_2, \dots, X_n)$ une suite de variables aléatoires discrètes, et soit (S_1, S_2, \dots, S_n) une suite d'ensembles finis ou dénombrables tels que, pour tout $i \leq n$, $\mathbb{P}(X_i \in S_i) = 1$. Alors (X_1, X_2, \dots, X_n) est une suite de variables aléatoires indépendantes si et seulement si, pour tout $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \prod_{i=1}^n S_i$,

$$\mathbb{P}(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^n \mathbb{P}(X_i = x_i).$$

3.4.2 Cas des variables aléatoires à densité

Soit une suite $\mathbf{X} = (X_1, X_2, \dots, X_n)$ de variables aléatoires réelles définies sur le même espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ et de densités de probabilité respectives f_i . Les variables aléatoires réelles (X_1, X_2, \dots, X_n) sont dites indépendantes si et seulement si le vecteur \mathbf{X} a une densité de probabilité f qui se définit par

$$\forall \mathbf{x} = (x_1, \dots, x_n)^T \in \mathbb{R}^n, \quad f(x_1, \dots, x_n) = \prod_{i=1}^n f_i(x_i).$$

3.5 Quelques statistique sur vecteurs aléatoires

Comme dans le cas de variables aléatoires, nous allons maintenant donner quelques statistiques qu'on peut en tirer à partir d'un échantillon de vecteurs aléatoires tirés d'une population.

Soit un échantillon de valeurs de vecteurs aléatoires continues $(\mathbf{x}_1, \dots, \mathbf{x}_n)$, $\mathbf{x}_i \in \mathbb{R}^d$, on a le vecteur moyenne empirique et la matrice de covariance empirique sont respectivement données par

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_i \mathbf{x}_i \quad (3.9)$$

$$\mathbf{S} = \frac{1}{n} \sum_i (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T. \quad (3.10)$$

Chapitre 4

Variables aléatoires gaussiennes et vecteurs aléatoires gaussiens

L'une des densités les plus répandues en probabilité et statistique est la *densité normale* dite aussi *densité gaussienne* par référence à celui qui l'a proposé C. F. Gauss. Elle sert à modéliser de nombreux phénomènes, et comme on le verra plus tard, la distribution de la moyenne d'un échantillon de variables aléatoires issues d'une densité quelconque tend vers une loi normale quand la taille de l'échantillon augmente. Ceci montre donc la grande importance de la loi normale.

4.1 Loi normale mono-dimensionnelle

4.1.1 Loi normale générale

Définition 4.1.1. On appelle loi ou densité normale (ou gaussienne) univariée de paramètres (μ, σ^2) (où $\sigma \geq 0$) la loi de probabilité définie par la densité $f : \mathbb{R} \rightarrow \mathbb{R}^+$, telle que pour tout $x \in \mathbb{R}$:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \quad (4.1)$$

Soit X une variable aléatoire réelle admettant pour densité de probabilité la loi normale (4.1), appelée variable aléatoire gaussienne, son espérance est μ et son écart type est σ (sa variance est σ^2). On note

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

et on lit “ X suit la loi normale d'espérance μ et de variance σ^2 ”.

Sa densité de probabilité dessine une courbe dite courbe en cloche ou courbe de Gauss.

4.1.2 Loi normale centrée réduite

La loi normale centrée réduite correspond à un cas particulier de la loi normale.

Définition 4.1.2. On appelle loi normale (ou gaussienne) centrée réduite la loi définie par la densité de probabilité $\varphi : \mathbb{R} \rightarrow \mathbb{R}^+$ définie par :

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad (4.2)$$

qui se note $\mathcal{N}(0, 1)$. Elle est dite centrée de part son espérance nulle et réduite de part sa variance unité.

4.1.3 Fonction de répartition

La fonction de répartition de la loi normale générale s'exprime communément à l'aide de la fonction caractéristique de la loi normale centrée réduite.

On note Φ la fonction de répartition de la loi normale centrée réduite. Elle est définie, pour tout réel x , par :

$$\Phi(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x \varphi(u) du = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du. \quad (4.3)$$

Cette primitive ne s'exprime pas à l'aide des fonctions usuelles (exponentielle, etc.) mais devient elle-même une fonction usuelle. La fonction Φ s'exprime à l'aide de la fonction d'erreur notée erf (error function) (appelée aussi fonction d'erreur de Gauss)

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-v^2} \quad (4.4)$$

par

$$\Phi(x) = \frac{1}{2} \left(1 + \operatorname{erf} \left(\frac{x}{\sqrt{2}} \right) \right). \quad (4.5)$$

△ Remarque 4.1.1. ou a encore

$$\operatorname{erf}(x) = 2\Phi(x\sqrt{2}) - 1. \quad (4.6)$$

La fonction de répartition F pour la loi normale générale est donc donnée par, pour tout $x \in \mathbb{R}$,

$$F(x) = \Phi \left(\frac{x - \mu}{\sigma} \right). \quad (4.7)$$

4.1.4 Quelques propriétés

Soit une variable aléatoire X qui suit la loi normale $\mathcal{N}(\mu, \sigma^2)$. Alors :

1. son espérance et sa variance existent et $\mathbb{E}(X) = \mu$, $\operatorname{Var}(X) = \sigma^2 \geq 0$,
2. la variable aléatoire $X^* = \frac{X - \mathbb{E}(X)}{\sqrt{\operatorname{Var}(X)}}$, c'est-à-dire $X^* = \frac{X - \mu}{\sigma}$, suit la loi normale centrée réduite $\mathcal{N}(0, 1)$,
3. si a, b sont deux réels ($a \neq 0$), alors la variable aléatoire $aX + b$ suit la loi normale $\mathcal{N}(a\mu + b, a^2\sigma^2)$
4. Si une v.a. X suit la loi normale $\mathcal{N}(\mu, \sigma^2)$, alors la v.a. $\exp(X)$ suit la loi log-normale.

5. Si les variables X_1, X_2, \dots, X_n sont indépendantes et de loi commune $\mathcal{N}(0, 1)$, alors la somme de leur carré : $\sum_{k=1}^n X_k^2$ suit une loi du χ^2 à n degrés de liberté : $\chi^2(n)$.
6. La formule s'étend pour des variables normales non centrées et non réduites : si les variables X_1, X_2, \dots, X_n sont indépendantes, chaque X_i et de loi $\mathcal{N}(\mu_i, \sigma_i^2)$, alors la somme de leurs carrés centrés réduits $\sum_{i=1}^k \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2$ suit une loi du χ^2 à n degrés de liberté : $\chi^2(n)$.
7. Si U et V sont deux v.a. indépendantes de loi uniforme sur $[0, 1]$, alors les deux v.a. $X = \sqrt{-2 \ln(U)} \cos(2\pi V)$ et $Y = \sqrt{-2 \ln(U)} \sin(2\pi V)$ sont de loi normale centrée réduite. De plus X et Y sont indépendantes. (Ces deux formules sont utilisées pour simuler la loi normale.)

4.2 Théorème central limite

La grande importance pratique associée à la distribution normale découle du théorème central limite présenté ci-dessous (théorème 4.2.1).

Théorème 4.2.1. Le Théorème Central Limite. Soit X_1, X_2, \dots, X_n une suite de variables aléatoires réelles mutuellement indépendantes et suivant la même densité f (variables i.i.d) d'espérance μ et d'écart-type σ . Soit la variable aléatoire somme

$$S = X_1 + X_2 + \dots + X_n$$

Alors la densité de probabilité de la somme S converge vers la loi normale $\mathcal{N}(n\mu, n\sigma^2)$ quand n tend vers l'infini :

$$\lim_{n \rightarrow \infty} f(s) = \mathcal{N}(n\mu, n\sigma^2)$$

Ceci montre donc l'importance que joue la loi normale pour approximer la densité de données issues de l'accumulation de plusieurs phénomènes notamment physiques.

⚠ Remarque 4.2.1. Ce théorème peut aussi se formuler ainsi. Soit la variable aléatoire moyenne

$$\bar{X} = \frac{S}{n} = \frac{X_1 + \dots + X_n}{n},$$

et la variable aléatoire centrée réduite

$$Z = \frac{S - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

On a donc la densité de Z converge vers la loi normale centrée réduite quand n tend vers l'infini.

4.3 Couple de variables aléatoires gaussiennes

Considérons deux variables aléatoires réelles X et Y . La densité jointe $f_{X,Y}$ du couple (X, Y) est dite normale (gaussienne) si elle prend la forme suivante :

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y(1-\rho^2)^{1/2}} \times \exp \left\{ -\frac{1}{2(1-\rho^2)} \left(\left(\frac{x-\mu_X}{\sigma_X} \right)^2 - 2\rho \frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y} \right)^2 \right) \right\}. \quad (4.8)$$

avec σ_1 et $\sigma_2 \geq 0$ et $\rho \in [-1, 1]$. On montre que $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ et $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ et que ρ est le coefficient de corrélation entre X et Y .

Théorème 4.3.1. *Une corrélation nulle (décorrélacion) implique l'indépendance quand les variables aléatoires sont gaussiennes.*

Démonstration. Soit $\rho = 0$ dans l'équation (4.8). On a la densité jointe de (X, Y) est donnée par

$$\begin{aligned} f_{X,Y}(x, y) &= \frac{1}{2\pi\sigma_X\sigma_Y} \exp \left\{ -\frac{1}{2} \left(\left(\frac{x-\mu_X}{\sigma_X} \right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y} \right)^2 \right) \right\}. \\ &= \frac{1}{\sqrt{2\pi}\sigma_X} \exp \left\{ -\frac{1}{2} \left(\frac{x-\mu_X}{\sigma_X} \right)^2 \right\} \times \frac{1}{\sqrt{2\pi}\sigma_Y} \exp \left\{ -\frac{1}{2} \left(\frac{y-\mu_Y}{\sigma_Y} \right)^2 \right\} \\ &= f_X(x)f_Y(y) \end{aligned} \quad (4.9)$$

□

qui est le résultat recherché.

4.4 Vecteurs aléatoires gaussiens

Définition 4.4.1. *Un vecteur aléatoire $\mathbf{X} = (X_1, \dots, X_d)^T$ est dit gaussien si, pour tout $\mathbf{u} \in \mathbb{R}^d$, la variable aléatoire réelle $\mathbf{u}^T \mathbf{X}$ est une variable aléatoire gaussienne. C'est à dire si toute combinaison linéaire de ses composantes est une variable gaussienne.*

4.4.1 Loi normale multidimensionnelle

La loi normale multidimensionnelle représente la distribution d'un vecteur aléatoire gaussien. Elle généralise donc la loi normale monodimensionnelle. Soit \mathbf{X} un vecteur aléatoire gaussien de dimension d . On appelle loi normale multidimensionnelle d'espérance $\boldsymbol{\mu}$ et de matrice de covariance $\boldsymbol{\Sigma}$ la densité de probabilité $f : \mathbb{R}^d \rightarrow \mathbb{R}^+$ définie par :

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right), \quad (4.10)$$

$|\boldsymbol{\Sigma}|$ étant le déterminant de $\boldsymbol{\Sigma}$. La loi normale multidimensionnelle se note habituellement $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

⚠ Remarque 4.4.1. Le terme présent dans l'exponentielle $(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ est le carré de la **distance de Mahalanobis**. Cette dernière est en effet donnée par $\sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}$

4.4.2 Quelques propriétés

Soit $\mathbf{X} = (X_1, \dots, X_d)^T$ un vecteur aléatoire gaussien de dimension d suivant la loi $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, on a :

1. l'espérance et la matrice de covariance de \mathbf{X} sont respectivement $\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu}$ et $\text{Cov}(\mathbf{X}) = \boldsymbol{\Sigma}$,
2. les variables aléatoires X_1, \dots, X_d sont indépendantes si et seulement si la matrice de covariance est diagonale,
3. Tout sous-vecteur d'un vecteur aléatoire gaussien suit une loi normale. En particulier, ses composantes sont toutes gaussiennes,
4. Si \mathbf{A} est une matrice constante de dimensions $[n \times d]$ et \mathbf{b} un vecteur constant dans \mathbb{R}^n , alors la densité du vecteur aléatoire $\mathbf{Y} = \mathbf{A} \mathbf{X} + \mathbf{b}$ (de dimension $[n \times 1]$) est la loi normale de n dimensions suivante $\mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$.

Chapitre 5

Estimation de paramètres

5.1 Introduction

Supposons que, pour étudier et caractériser un phénomène physique, naturel ou autre, nous avons choisi et adopté un modèle probabiliste paramétrique représenté par une fonction de densité de probabilité $f(x; \theta)$ (ou une fonction de masse de probabilité $P(x; \theta)$ dans le cas discret). L'explication de ce phénomène nécessite donc l'estimation de ce modèle probabiliste à partir des données que l'on a observées (l'échantillon que l'on a à notre disposition). Ceci consiste donc à estimer le(s) paramètre(s) θ de ce modèle à partir des données observées (x_1, \dots, x_n) que l'on va supposer indépendantes dans le cadre de ce cours. Nous considérerons d'abord le cas d'un seul paramètre θ à estimer pour plus de clarté et simplicité et notons par $f(x; \theta)$ la densité ayant θ comme vrai paramètre mais qui est inconnu et que l'on cherche à estimer.

Le problème d'estimation de paramètres est donc celui de déterminer une fonction appropriée des données (x_1, \dots, x_n) , que nous noterons $h(x_1, \dots, x_n)$ qui donne la "meilleure" estimation de θ au sens de critère d'optimalité que nous verrons.

Nous avons donc

$$\hat{\theta} = h(x_1, \dots, x_n)$$

et plus généralement, sous forme de variable aléatoire (car en effet pour des nouvelles réalisations des X_j , la valeur de $\hat{\theta}$ change) :

$$\hat{\Theta} = h(X_1, \dots, X_n).$$

Cette statistique à déterminer s'appelle *un estimateur*.

5.2 Critères de qualité pour les estimations

Nous verrons maintenant un certain nombre de critères selon lesquels la qualité d'une estimation peut être évaluée. Ces critères définissent en général des propriétés souhaitables pour un estimateur et fournissent un guide par lequel la qualité d'un estimateur peut être comparée à celle d'un autre.

Notre objectif est de déterminer une statistique

$$\hat{\Theta} = h(X_1, \dots, X_n)$$

qui fournit une bonne estimation de θ . Cette statistique à déterminer s'appelle *un estimateur* de θ , pour lequel des propriétés comme la moyenne, la variance ou la distribution fournissent une mesure de qualité pour cet estimateur. Une fois nous avons observé un échantillon de valeurs (x_1, \dots, x_n) la valeur de l'estimateur

$$\hat{\theta} = h(x_1, \dots, x_n)$$

qui est une valeur numérique, est appelé *estimation* du paramètre θ .

Propriété 5.2.1. Absence de biais. *Un estimateur $\hat{\Theta}$ de θ est dit sans biais si*

$$\mathbb{E}[\hat{\Theta}] = \theta, \quad (5.1)$$

Ceci est clairement une propriété désirée pour $\hat{\Theta}$, qui consiste à dire que, en moyenne, on espère que $\hat{\Theta}$ est égal à la valeur du vrai paramètre θ .

Il est naturel que, si $\hat{\Theta} = h(X_1, \dots, X_n)$ est à qualifier comme un bon estimateur de θ , non seulement sa moyenne doit être très proche du vrai paramètre θ mais aussi il faudrait qu'il y ait une grande probabilité que toute valeur $\hat{\theta}$ soit très proche de θ . Cela revient à sélectionner un estimateur de façon à ce que non seulement il soit sans biais mais aussi sa variance soit la plus petite possible. Ainsi, la seconde qualité désiré d'un estimateur et celle de *variance minimale*.

Propriété 5.2.2. Variance minimale. *Soit $\hat{\Theta}$ un estimateur sans biais de θ . Il est dit à variance minimale pour θ si, pour tout autre estimateur sans biais Θ^* de θ , à partir du même échantillon, on a :*

$$\text{var}(\hat{\Theta}) \leq \text{var}(\Theta^*). \quad (5.2)$$

Étant donné deux estimateurs sans biais pour un paramètre donné, celui ayant la variance plus faible est préférable, car une plus petite variance implique que les valeurs observées de l'estimateur (les estimations) ont tendance à être plus proche de sa moyenne qui est la valeur du vrai paramètre.

La question qui se pose donc est, étant donné un échantillon à partir duquel on construit plusieurs estimateurs sans biais, le quel parmi tout ces estimateurs qui a la variance minimale. Cette question est difficile mais, il existe un théorème qui montre qu'il est possible de déterminer la variance minimale possible (borne inférieure sur la variance) de tout estimateur sans biais obtenu à partir d'un échantillon donné.

Théorème 5.2.1. Borne de Cramer-Rao *Soit (X_1, \dots, X_n) un échantillon de v.a issues d'une population de densité $f(x; \theta)$ où θ est le paramètre inconnu, et soit $\hat{\Theta} = h(X_1, \dots, X_n)$ un estimateur sans biais pour θ . La variance de $\hat{\Theta}$ satisfait l'inégalité suivante*

$$\text{var}(\hat{\Theta}) \geq \left[n \mathbb{E} \left(\frac{\partial \ln f(X; \theta)}{\partial \theta} \right)^2 \right]^{-1} \quad (5.3)$$

si l'espérance et la dérivée existent. Un résultat analogue avec $p(X; \theta)$ en remplaçant $f(X; \theta)$ est obtenue lorsque X est discrète. Cette inéquation fournit donc une borne inférieure de la variance de n'importe quel estimateur sans biais. On note aussi que cette borne s'exprime généralement en fonction du vrai paramètre θ .

La quantité $n \mathbb{E} \left(\frac{\partial \ln f(X; \theta)}{\partial \theta} \right)^2$ s'appelle l'*information de Fisher* contenue dans un échantillon de taille n et se note $\mathcal{I}_n(\theta)$. La borne inférieure de Cramér-Rao alors se définit aussi par :

$$\text{var}(\hat{\Theta}) \geq \frac{1}{\mathcal{I}_n(\theta)} \quad (5.4)$$

et énonce donc que l'inverse de l'information de Fisher, $\mathcal{I}_n(\theta)$, d'un paramètre θ , est une borne inférieure de la variance d'un estimateur sans biais de ce paramètre.

⚠ Remarque 5.2.1. *En anglais, la borne inférieure de Cramér-Rao s'appelle **Cramér-Rao Lower Bound** abrégée par CRLB.*

⚠ Remarque 5.2.2. Deuxième forme opérationnelle. *Si le modèle est régulier, l'espérance $\left[\mathbb{E} \left(\frac{\partial \ln f(X; \theta)}{\partial \theta} \right)^2 \right]^{-1}$ dans (5.3) est équivalente à $-\left[\mathbb{E} \left(\frac{\partial^2 \ln f(X; \theta)}{\partial^2 \theta} \right) \right]^{-1}$. L'inégalité de Cramér-Rao peut également être alors mise sous la deuxième forme opérationnelle :*

$$\text{var}(\hat{\Theta}) \geq - \left[n \mathbb{E} \left(\frac{\partial^2 \ln f(X; \theta)}{\partial^2 \theta} \right) \right]^{-1}. \quad (5.5)$$

Cette expression alternative souvent offre des avantages de point de vue calcul.

5.2.1 Cas de plusieurs paramètres

Le résultat donné par l'inéquation (5.3) peut être facilement étendu au cas de plusieurs paramètres. Soient $(\theta_1, \dots, \theta_m)$ ($m \leq n$) les paramètres inconnus du modèle (la densité) $f(x; \theta_1, \dots, \theta_m)$ que l'on cherche à estimer à partir d'un échantillon de données de taille n . Utilisons la notation vectorielle suivante pour le vecteur paramètre et le vecteur pour les estimateurs :

$$\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)^T$$

et

$$\hat{\boldsymbol{\Theta}} = (\hat{\Theta}_1, \dots, \hat{\Theta}_m)^T.$$

De façon similaire à (5.3), on peut montrer que l'inégalité de Cramér-Rao, pour le cas de paramètres multiples, est de la forme

$$\text{cov}(\hat{\boldsymbol{\Theta}}) \geq \frac{\Lambda^{-1}}{n}, \quad (5.6)$$

ou le terme général de la matrice Λ est donné par :

$$\Lambda_{ij} = \Lambda(\theta_i, \theta_j) = \mathbb{E} \left[\left(\frac{\partial \ln f(X; \boldsymbol{\theta})}{\partial \theta_i} \right) \left(\frac{\partial \ln f(X; \boldsymbol{\theta})}{\partial \theta_j} \right) \right], \quad i, j = 1, 2, \dots, m. \quad (5.7)$$

On a donc remplacé l'information de Fisher par la matrice Λ qui est la *matrice d'information de Fisher*.

Pour le j ème paramètre, l'inégalité (5.7) implique que

$$\text{var}(\hat{\Theta}_j) \geq \frac{1}{n \Lambda_{jj}}, \quad j = 1, \dots, m.$$

⚠ Remarque 5.2.3. *La CRLB peut être transformée sous une transformation du paramètre. Supposons que, au lieu de θ , on s'intéresse à $\phi = g(\theta)$ qui est une transformation un-à-un et différentiable par rapport à θ ; alors*

$$\text{CRLB pour var}(\hat{\boldsymbol{\Phi}}) = \left[\frac{d g(\theta)}{d \theta} \right]^2 \times \left(\text{CRLB pour var}(\hat{\boldsymbol{\Theta}}) \right) \quad (5.8)$$

où $\hat{\boldsymbol{\Phi}}$ est un estimateur sans biais pour ϕ .

⚠ **Remarque 5.2.4. Efficacité d'un estimateur.** Étant donné un estimateur sans biais $\hat{\Theta}$ de θ , le rapport de sa CRLB par sa variance est appelé l'**efficacité** de $\hat{\Theta}$

$$\begin{aligned} e(\hat{\Theta}) &= \frac{\text{CRLB pour } \text{var}(\hat{\Phi})}{\text{var}(\hat{\Theta})} \\ &= \cdot \end{aligned} \quad (5.9)$$

L'efficacité d'un estimateur sans biais est ainsi inférieure ou égale à 1. Un estimateur sans biais ayant une efficacité égale à 1 est dit **efficace**.

Pour un estimateur, on souhaite aussi pouvoir, en augmentant la taille de l'échantillon, diminuer l'erreur d'estimation. Si c'est le cas, on dit que l'estimateur est consistant (on dit aussi convergent).

Propriété 5.2.3. Consistance (ou convergence). Un estimateur $\hat{\Theta}$ est dit **consistant** pour θ si,

$$\lim_{n \rightarrow \infty} \mathbb{P}(|\hat{\Theta} - \theta| > \epsilon) = 0 \quad \forall \epsilon > 0. \quad (5.10)$$

On l'interprète comme le fait que la probabilité de s'éloigner de la vraie valeur du paramètre de plus de ϵ tend vers 0 quand la taille de l'échantillon augmente. L'estimateur donc converge vers la valeur à estimer quand la taille de l'échantillon augmente.

Théorème 5.2.2. Soit $\hat{\Theta}$ un estimateur pour θ sur un échantillon de taille n . Alors, si

$$\lim_{n \rightarrow \infty} \mathbb{E}[\hat{\Theta}] = \theta, \quad \text{et} \quad \lim_{n \rightarrow \infty} \text{var}[\hat{\Theta}] = 0, \quad (5.11)$$

l'estimateur $\hat{\Theta}$ est dit *consistant* pour θ .

Afin de décrire une autre propriété d'un estimateur, qui est la *suffisance*, on introduit d'abord la notion de statistique suffisante.

Définition 5.2.1. Soit X un vecteur d'observations de taille n avec les X_i i.i.d.. Soit θ un paramètre de la loi de probabilité des X_i . Une statistique $S(X)$ est dite *exhaustive* pour le paramètre θ (on dit aussi *suffisante*) si la probabilité conditionnelle d'observer X sachant $S(X)$ est indépendante de θ . Cela peut se traduire par la formule suivante :

$$\mathbb{P}(X = x | S(X) = s, \theta) = \mathbb{P}(X = x | S(X) = s), \quad (5.12)$$

En pratique l'on se sert peu de cette formule pour montrer qu'une statistique est exhaustive et l'on préfère en règle générale utiliser le critère suivant appelé *critère de factorisation* (parfois aussi appelé *critère de Fisher-Neyman*) :

Soit $f_\theta(x)$ la densité de probabilité du vecteur aléatoire X . Une statistique S est exhaustive si et seulement s'il existe deux fonctions g et h telles que : $f_\theta(x) = h(x)g(\theta, S(x))$,

Propriété 5.2.4. Estimateur suffisant. Soit (X_1, X_2, \dots, X_n) un échantillon i.i.d. de X de distribution à paramètre θ . Si $Y = h(X_1, X_2, \dots, X_n)$ est une statistique telle que, pour toute autre statistique $Z = g(X_1, X_2, \dots, X_n)$, la distribution conditionnelle de Z , étant donné $Y = y$, ne dépend pas de θ , ç.à.d

$$\mathbb{P}(Z = z | Y = y, \theta) = \mathbb{P}(Z = z | Y = y)$$

, alors Y est appelée une **statistique exhaustive (suffisante)** pour θ . Si l'on a également $\mathbb{E}[Y] = \theta$, alors Y est dit un **estimateur suffisant** pour θ .

Autrement dit, la définition de la suffisance dit que, si Y est une statistique suffisante pour θ , toute l'information de l'échantillon concernant θ est contenue dans Y .

Si une statistique suffisante pour un paramètre θ existe, le théorème 5.2.3 suivant, fournit un moyen de la trouver.

Théorème 5.2.3. Critère de factorisation de Fisher-Neyman Soit $Y = h(X_1, \dots, X_n)$ une statistique basée sur un échantillon i.i.d de taille n . Alors Y est une statistique exhaustive pour θ si et seulement si la densité jointe des X_i peut être factorisée selon la forme :

$$f(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta) = g(h(x_1, \dots, x_n), \theta) \phi(x_1, \dots, x_n). \quad (5.13)$$

Dans le cas discret on a :

$$P(x_1, \dots, x_n; \theta) = \prod_{i=1}^n P(x_i; \theta) = g(h(x_1, \dots, x_n), \theta) \phi(x_1, \dots, x_n). \quad (5.14)$$

Le résultat ci-dessus peut être étendu au cas de paramètres multiples. Soit $\theta = (\theta_1, \dots, \theta_m)^T$, $m \leq n$ le vecteur paramètre. Alors $Y_1 = h(X_1, \dots, X_n), \dots, Y_r = h(X_1, \dots, X_n)$, $r \geq m$ est un ensemble de statistique suffisantes pour θ si et seulement si

$$f(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta) = g(\mathbf{h}(x_1, \dots, x_n), \theta) \phi(x_1, \dots, x_n). \quad (5.15)$$

avec $\mathbf{h} = (h_1, \dots, h_n)^T$. L'expression dans le cas discret est similaire.

5.3 Méthodes d'estimation

Il existe plusieurs méthodes d'estimation de paramètres, notamment estimation ponctuelle comme la méthode des moments, la méthode du maximum de vraisemblance, la méthode du maximum a posteriori ou la méthode d'estimation par intervalle. Dans ce cours, nous verrons en particulier les méthodes du maximum de vraisemblance et de maximum a posteriori (estimation ponctuelle) pour leur usage très commun en estimation de modèles probabilistes et aussi la méthode d'estimation par intervalle de confiance.

L'estimation d'un paramètre quelconque θ est *ponctuelle* si l'on associe une seule valeur à l'estimateur à partir des données observées sur un échantillon aléatoire. L'estimation *par intervalle* associe quant à elle à un échantillon aléatoire, un intervalle $[\hat{\theta}_a, \hat{\theta}_b]$ qui recouvre θ avec une certaine probabilité.

Chapitre 6

Méthode du maximum de vraisemblance

Introduite par le statisticien Fischer en 1922, la méthode du maximum de vraisemblance est devenue la méthode générale la plus importante de l'estimation d'un point de vue théorique. Son plus grand atout réside dans le fait que certaines propriétés très générale associées à cette procédure peuvent être dérivées et, dans le cas de grands échantillons, ce sont des propriétés optimales en fonction des critères d'absence de biais, de variance minimale, de consistance et d'efficacité.

6.1 Définition de la fonction de vraisemblance

Soit $f(x; \theta)$ la fonction de densité de probabilité d'une a.a X où θ est le paramètre (vrai paramètre) à estimer (Nous prenons ici le cas simple d'un seul paramètre). Soit $\mathbf{x} = (x_1, \dots, x_n)$ un échantillon d'observations des variables aléatoires (X_1, \dots, X_n) . La *vraisemblance* du paramètre θ pour l'échantillon \mathbf{x} est donnée par la densité jointe de \mathbf{x} et se note ainsi :

$$L(\theta; x_1, \dots, x_n) = L(\theta; \mathbf{x}) = f(x_1, \dots, x_n; \theta). \quad (6.1)$$

Sous l'hypothèse que les individus son *i.i.d.*, on a donc

$$\begin{aligned} L(\theta; x_1, \dots, x_n) &= f(x_1; \theta) f(x_2; \theta) \cdots f(x_n; \theta) \\ &= \prod_{i=1}^n f(x_i; \theta). \end{aligned} \quad (6.2)$$

Dans le cas où les X_i sont des v.a discrètes, on a

$$\begin{aligned} L(\theta; x_1, \dots, x_n) = L(\theta; \mathbf{x}) &= P(x_1; \theta) P(x_2; \theta) \cdots P(x_n; \theta) \\ &= \prod_{i=1}^n P(x_i; \theta). \end{aligned} \quad (6.3)$$

⚠ Remarque 6.1.1. On peut aussi rencontrer la notation $L(\theta)$ de la vraisemblance de θ au lieu de $L(\theta; x_1, \dots, x_n)$.

On voit que pour des valeurs d'échantillon données, la fonction de vraisemblance est seulement fonction du paramètre θ .

6.2 Maximum de vraisemblance

Définition 6.2.1. Maximum de vraisemblance. *L'estimation de θ par la méthode du maximum de vraisemblance consiste à choisir, comme estimation de θ , la valeur de θ qui maximise la fonction de vraisemblance $L(\theta)$.*

En effet, en choisissant une valeur de θ qui maximise L (ou $\ln L$), cela revient à dire que, parmi les valeurs possible de θ , nous prenons la valeur qui rend le plus probable que possible l'évènement que les les valeurs de l'échantillon observé (x_1, \dots, x_n) viennent de la population de densité $f(x; \theta)$.

6.2.1 Cas d'un seul paramètre à estimer

Mathématiquement, le maximum de $L(\theta)$ correspond à la valeur de θ pour laquelle la dérivée de L par rapport à θ est nulle : $\frac{dL(\theta)}{d\theta} = 0$ (extremum) et la dérivée seconde par rapport à θ est négative : $\frac{d^2L(\theta)}{d^2\theta} < 0$ pour identifier le maximum parmi les possibles extrema obtenus. Ainsi, l'*estimateur du maximum de vraisemblance* (MV) de θ , souvent noté $\hat{\theta}$, à partir des valeurs de l'échantillon (x_1, \dots, x_n) peut être déterminé à partir de

$$\left. \frac{dL(\theta; x_1, \dots, x_n)}{d\theta} \right|_{\theta=\hat{\theta}} = 0. \quad (6.4)$$

qui permet d'identifier les extrema de $L(\theta)$ (mais ne permet pas de savoir lesquels parmi ces extrema sont des maxima (que nous recherchons) ou bien des minima (qui ne nous intéressent pas). Il faut donc, après que les solutions de l'équation aient été trouvées, sélectionner celles qui correspondent à des maxima. Un authentique vérifie

$$\left. \frac{d^2L(\theta; x_1, \dots, x_n)}{d^2\theta} \right|_{\theta=\hat{\theta}} < 0 \quad (6.5)$$

il faut donc sélectionner parmi les solutions de la première équation celles qui vérifient cette deuxième équation.

Bien que la plupart des vraisemblances soient différentiables (avec l'importante exception de la distribution uniforme), les solutions de l'équation (6.4) ne s'expriment pas toujours par des formes analytiques. On a souvent recours à des méthodes d'optimisations numériques pour identifier les maxima de la fonction de vraisemblance (par exemple comme en régression Logistique, mélange de densités, modèles de Markov cachés, etc) comme par exemple la montée de gradient, l'algorithme de Newton Raphson, l'algorithme EM, etc.

De manière équivalente, cela revient à annuler la dérivée du logarithme de la fonction de vraisemblance, la fonction de vraisemblance étant en effet positive et le logarithme est monotone et la vraisemblance atteint donc son maximum pour la même valeur que son logarithme. Le logarithme de la fonction de vraisemblance s'appelle *log-vraisemblance*. Manipuler le log de la fonction de vraisemblance au lieu de la vraisemblance elle même vient aussi du fait que, comme cette dernière s'écrit souvent comme produit de densités (de probabilités dans le cas discrets), cela peut résulter en des valeurs très faibles qui peuvent dans certains cas dépasser la précision de calculateurs. Ainsi, traiter des logarithmes revient plutôt à sommer et donc d'éviter des problèmes numériques.

L'estimateur de MV de θ est donc aussi donné par

$$\frac{d \ln L(\theta; x_1, \dots, x_n)}{d \theta} \Big|_{\theta=\hat{\theta}} = 0. \quad (6.6)$$

C'est l'équation de vraisemblance.

La solution désirée est une racine de cette equation, ou par équivalence celle en L , (si cette fonction admet des racines). Dans le cas où cette fonction est concave et admet donc une seule racine, l'estimateur du maximum de vraisemblance correspond à cette racine et on parle de *maximum global*. Cependant, la fonction de vraisemblance peut avoir plus d'un maximum (maxima). Dans ce cas, on parle de *maxima locaux* et l'estimateur du maximum de vraisemblance correspond au maximum global (lorsque tous les maxima ont été identifiés, seul le plus grand d'entre-eux doit être retenu).

6.2.2 Cas de plusieurs paramètres à estimer (Vraisemblance multivariée)

Plusieurs densités admettent plus d'un paramètre. Par exemple l'estimation d'une densité normale monodimensionnelle nécessite l'estimation de la moyenne μ et de la variance σ^2 . L'extension au cas de plusieurs paramètres est simple. Dans le cas de m paramètres à estimer, la fonction de vraisemblance devient

$$\ln L(\theta_1, \dots, \theta_m; x_1, \dots, x_n)$$

et les estimateurs de MV de θ_j , $j = 1, \dots, m$, sont obtenus en résolvant simultanément le système d'équations de vraisemblance

$$\frac{d \ln L(\theta_1, \dots, \theta_m; x_1, \dots, x_n)}{d \theta_j} \Big|_{\theta_j=\hat{\theta}_j} = 0 \quad \text{pour } j = 1, \dots, m \quad (6.7)$$

et comme pour le cas univarié, mais dans ce cas multivarié c'est plus complexe, il faut en plus que au moins une des dérivées partielles secondes de L soit strictement négative pour au moins un j

$$\frac{d^2 \ln L(\theta_1, \dots, \theta_m; x_1, \dots, x_n)}{d^2 \theta_j} \Big|_{\theta_j=\hat{\theta}_j} = 0 \quad \text{pour au moins un } j$$

et le déterminant de la matrice des dérivées partielles secondes de L soit strictement positif :

$$\left| \frac{d^2 \ln L(\theta_1, \dots, \theta_m; x_1, \dots, x_n)}{d^2 \theta_j} \right|_{\theta_j=\hat{\theta}_j} > 0$$

Cette dernière condition est en général difficile à vérifier, même dans les cas simples.

6.3 Propriétés

Soit $\hat{\theta}$ la valeur de l'estimateur du maximum de vraisemblance de θ estimée à partir de l'échantillon (x_1, \dots, x_n) de la population X de densité $f(x; \theta)$ L'estimation MV $\hat{\theta}$, calculée à partir de (x_1, \dots, x_n) peut être notée comme

$$\hat{\theta} = h(x_1, \dots, x_n).$$

L'estimateur du maximum de vraisemblance $\hat{\Theta}^1$ pour θ est donc

$$\hat{\Theta} = h(X_1, \dots, X_n).$$

Propriété 6.3.1. Consistance. L'estimateur obtenu par la méthode du maximum de vraisemblance est convergent : $\lim_{n \rightarrow \infty} \mathbb{P}(|\hat{\Theta}_n - \theta| > \epsilon) = 0 \quad \forall \epsilon > 0.$

Propriété 6.3.2. Absence de biais et efficacité asymptotiques. Soit $\hat{\Theta}$ l'estimateur du maximum de vraisemblance (EMV) pour θ sous la densité $f(x; \theta)$ à partir d'un échantillon de taille n . Alors, quand n tend vers l'infini on a :

$$\lim_{n \rightarrow \infty} \mathbb{E}[\hat{\Theta}] = \theta \quad (6.8)$$

L'EMV est donc asymptotiquement sans biais. On a également

$$\lim_{n \rightarrow \infty} \text{var}[\hat{\Theta}] = \frac{1}{n \mathbb{E} \left[\left(\frac{\partial f(X; \theta)}{\partial \theta} \right)^2 \right]} = \frac{1}{\mathcal{I}_n(\theta)} = \text{CRLB}. \quad (6.9)$$

L'EMV est donc asymptotiquement efficace. Des résultats analogues sont obtenus lorsque la loi de X est discrète.

Propriété 6.3.3. Normalité asymptotique En outre, la distribution de $\hat{\Theta}$ tend vers une distribution normale lorsque n devient grand. L'EMV est donc asymptotiquement normal.

$$\sqrt{n}(\hat{\Theta} - \theta) \xrightarrow[n \rightarrow \infty]{\text{loi}} \mathcal{N}(0, \mathcal{I}_n(\theta)^{-1}). \quad (6.10)$$

Propriété 6.3.4. Invariance. On peut montrer que, si $\hat{\Theta}$ est l'EMV de θ , alors l'EMV d'une fonction bijective différentiable de θ , soit $g(\theta)$, est $g(\hat{\Theta})$.

Cette importante propriété d'invariance implique que, par exemple, si $\hat{\Sigma}$ est l'EMV de l'écart type σ pour une distribution donnée, alors l'EMV de la variance σ^2 est $\hat{\Sigma}^2$.

6.4 Cas gaussien

Soit (X_1, \dots, X_n) un échantillon de variables aléatoires réelles issues d'une population de densité normale $\mathcal{N}(\mu, \sigma^2)$, alors la moyenne empirique $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ est un estimateur sans biais de l'espérance μ et la variance empirique corrigée $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ est un estimateur sans biais de la variance σ^2 et on a

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad (6.11)$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2 \quad (6.12)$$

On peut donc remarquer que $\frac{\sqrt{n}(\bar{X} - \mu)}{S}$ suit une loi de student de paramètre $n-1$. Cela vient du fait que $(\bar{X} - \mu) \sim \mathcal{N}(0, \frac{\sigma^2}{n})$ donc $\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim \mathcal{N}(0, 1)$ et comme S^2 suit une loi de χ^2 , en remplaçant σ par son estimateur S on a alors la loi de student t_{n-1} .

1. $\hat{\Theta}$ représente la variable aléatoire associée à $\hat{\theta}$.

6.5 Estimation par intervalle

Nous allons maintenant voir une autre approche d'estimation des paramètres l'estimation par intervalle. L'estimation par intervalle fournit, à partir d'un échantillon d'une population, non seulement des valeurs des paramètres à estimer, mais un intervalle de valeurs centré sur la valeur numérique estimée du paramètre inconnu avec un niveau de confiance donné. Ce niveau de confiance représente la probabilité que le vrai paramètre se trouve dans l'intervalle que l'on donne comme estimation. L'intervalle est appelé *intervalle de confiance* le niveau de confiance est aussi appelé *précision* ou *coefficient de confiance*.

6.6 Intervalle de confiance

Définition 6.6.1. Intervalle de confiance. Soit (X_1, \dots, X_n) un échantillon de variables aléatoires issues d'une population de densité $f(x; \theta)$, θ étant le (vecteur) paramètre à estimer. Supposons aussi que $T_1(X_1, \dots, X_n)$ et $T_2(X_1, \dots, X_n)$ deux statistiques sur l'échantillon telle que $T_1 < T_2$. L'intervalle $[T_1, T_2]$ est dit *intervalle de confiance* à $100(1 - \alpha)\%$ pour θ si

$$P[T_1 < \theta < T_2] = 1 - \alpha. \quad (6.13)$$

α représente le *risque* que le vrai paramètre θ ne soit pas dans cet intervalle et $1 - \alpha$ s'appelle niveau ou (coefficient) de confiance. Une estimation par intervalle de confiance sera donc d'autant meilleure que l'intervalle sera petit pour un coefficient de confiance grand (proche de 1) ou de manière équivalente pour un risque α proche de zéro. Les valeurs généralement prises pour $1 - \alpha$ sont 0.90, 0.95, 0.99, and 0.999. Les limites de l'intervalle T_1 et T_2 sont appelés respectivement la *limite inférieure de confiance* et *limite supérieure de confiance*.

6.7 Cas d'estimation d'une gaussienne

6.7.1 Intervalle de confiance pour μ dans $\mathcal{N}(\mu, \sigma^2)$ avec σ connu

Chapitre 7

Méthode des Moindres Carrées

7.1 Méthode des Moindres Carrées

La méthode des moindres carrés consiste à estimer les paramètres d'un modèle en minimisant les écarts quadratiques entre les données observées, d'une part, et leurs valeurs attendues, d'autre part

Très utilisée notamment en régression où l'on cherche à expliquer la variation d'une variable de sortie (expliquée) Y , par la variation d'une variable d'entrée (explicative, covariable) X

Compte tenu de la valeur de X , la meilleure prédiction de Y (en termes d'erreur quadratique) est l'espérance $f(X)$ de Y sachant X .

On dit que Y est une fonction de X plus un bruit (erreur) :

$$Y = f(X) + E \quad (7.1)$$

f est appelée la fonction de régression, et E est un bruit souvent supposé d'espérance nulle.

L'estimateur des MC à des propriétés optimales d'absence de biais, de variance minimale (sous certaines conditions)

7.1.1 Définition des Moindres Carrés

Critères des moindres carrés

Soit le modèle

$$Y_i = f(X_i) + E_i \quad (7.2)$$

La fonction f est à estimer à partir d'un échantillon des couples de covariables X_i et leur réponses $Y_i : ((x_1, y_1), \dots, (x_n, y_n))$

Cette estimation est effectuée en minimisant la somme des écarts (erreurs) quadratiques

Définition 7.1.1. *Définition : Critères des moindres carrés* L'erreur quadratique est donnée par la somme des carrés des résidus (Residual Sum of Squares (RSS)) :

$$\text{RSS} = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - f(x_i))^2. \quad (7.3)$$

7.1.2 Moindres Carrés

Erreur quadratique dans le cas d'une fonction paramétrique

Soit $f(x; \theta)$ une fonction de paramètre θ à estimer. La somme des écarts quadratique dans ce cas est donnée par

$$RSS(\theta) = \sum_{i=1}^n (y_i - f(x_i; \theta))^2. \quad (7.4)$$

Définition 7.1.2. Définition de l'estimateur des moindres carrés *L'estimation de θ par la méthode des moindres carrés consiste à choisir, comme estimation de θ , la valeur de θ qui minimise la fonction $RSS(\theta)$.*

$$\hat{\theta} = \arg \min_{\theta} RSS(\theta) \quad (7.5)$$

En effet, en choisissant une valeur de θ qui minimise $RSS(\theta)$, cela revient à dire que, parmi les valeurs possible de θ , nous prenons la valeur qui correspond à une erreur minimale que les réponses y s'écartent de $f(x; \theta)$ pour l'échantillon observé $((x_1, y_1), \dots, (x_n, y_n))$.

Cas d'un seul paramètre à estimer

Cas d'un seul paramètre θ

Définition 7.1.3. Estimateur des moindres carrés (MC) *L'estimateur de moindres carrés (MC) de θ , noté $\hat{\theta}$, à partir des valeurs de l'échantillon $((x_1, y_1), \dots, (x_n, y_n))$ peut être déterminé à partir de*

$$\frac{d RSS(\theta)}{d \theta} \Big|_{\theta=\hat{\theta}} = 0. \quad (7.6)$$

qui permet d'identifier les extrema de $RSS(\theta)$. Il faut donc, après que les solutions de l'équation aient été trouvées, sélectionner celles qui correspondent à des minima. Un minimum vérifie

$$\frac{d^2 RSS(\theta)}{d^2 \theta} \Big|_{\theta=\hat{\theta}} > 0 \quad (7.7)$$

il faut donc sélectionner parmi les solutions de la première équation celles qui vérifient cette deuxième équation.

Moindres Carrés

⚠ Remarque 7.1.1. *Bien que la plupart des critères d'EQ soient différentiables, la minimisation du critère des MC ne s'effectue pas toujours de façon analytique*

⇒ On a souvent recours à des méthodes d'optimisations numériques (par exemple comme en réseau de neurones, etc)

⇒ la descente de gradient, l'algorithme de Newton Raphson, etc

Dans le cas où la fonction d'erreur est convexe, l'estimateur du Moindres Carrés fournit le minimum global. Cependant, dans beaucoup de problèmes réels, la fonction d'erreur n'est pas convexe et l'on a un minimum local ; atteindre le minimum global n'est pas toujours garanti

Des procédures algorithmiques existent (plusieurs initialisations, etc) et peuvent permettre d'atteindre un "bon" minimum local

Cas de plusieurs paramètres à estimer

Dans le cas d'un paramètre multiple $\theta = (\theta_1, \dots, \theta_m)$, le critère d'erreur est donné par

$$\text{RSS}(\theta) = \text{RSS}(\theta_1, \dots, \theta_m)$$

Les estimateurs de MC de $\theta_j, j = 1, \dots, m$, sont obtenus en résolvant simultanément le système d'équations suivant

$$\frac{\partial \text{RSS}(\theta)}{\partial \theta_j} \Big|_{\theta_j = \hat{\theta}_j} = 0 \quad \text{pour } j = 1, \dots, m \quad (7.8)$$

7.1.3 Propriétés de l'estimateur des moindres carrés

Soit $\hat{\theta}$ la valeur de l'estimateur des Moindres Carrés $\hat{\Theta}$ de θ estimée à partir de l'échantillon $((x_1, y_1), \dots, (x_n, y_n))$ de taille n

Propriété 7.1.1. *Absence de biais* Si l'on suppose que les erreurs (le bruit) sont d'espérance nulle ($\mathbb{E}[E_i] = 0$), l'estimateur des MC est sans biais

Propriété 7.1.2. *variance minimale* Si les erreurs sont d'espérance nulle ($\mathbb{E}[E_i] = 0$) et homoscédastiques décorrélées ($\mathbb{E}[E_i^T E_i] = \sigma^2 \mathbf{I}$) L'EMC est alors à variance minimale

\Rightarrow efficace et est donc le meilleur estimateur sans biais

Ces propriétés sont valables quelle que soit la distribution des erreurs.

Si en plus on fait l'hypothèse de normalité sur les erreurs ($e_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$) :

Propriété 7.1.3. *Normalité* La distribution de $\hat{\Theta}$ est normale centrée sur le vrai paramètre θ

Chapitre 8

Régression linéaire

8.1 Introduction

Les outils développés dans les chapitres 8 et 6 précédents pour l'estimation de paramètres et l'évaluation de la qualité d'un estimateur seront appliqués dans ce chapitre à une famille de problèmes très connues en statistique et estimation qui est celle de la régression simple et multiple.

Une situation qui arrive couramment est celle dans laquelle une variable aléatoire Y est fonction d'une ou plusieurs variables indépendantes déterministes (x_1, \dots, x_m) . Par exemple le prix d'un logement (Y) est une fonction de sa localisation (x_1) et de son âge (x_2); la durée de vie d'un composant électronique (Y) peut être liée à la température (x_1), la pression (x_2), etc; la vitesse d'un automobiliste (Y) en fonction du temps t , etc. Notons que les variables indépendantes est aussi appelées *variables explicatives* car à travers elles on cherche à expliquer les variables Y qui sont dites *expliquées*¹

L'objectif est donc d'estimer "la relation" entre Y et les variables indépendantes (x_1, \dots, x_m) étant donné un échantillon (Y_i, \dots, Y_n) de la variable Y et les valeurs associées des variables explicatives $x_j, j = 1, \dots, m$ pour chaque valeurs observée de $Y_i, i = 1 \dots, n$.

8.2 Le modèle linéaire simple

prenons le cas simple où l'on suppose que Y ne dépend que d'une seule variable explicative x et que cette relation est supposée linéaire. en d'autres termes on a la relation suivante

$$Y = \beta_0 + \beta_1 x + \epsilon, \quad (8.1)$$

avec $(\beta_0, \beta_1) \in \mathbb{R}^2$ sont les deux paramètres de la droite de régression, β_0 appelé *ordonnée à l'origine (intercept)* et β_1 *pente (slope)*. Ce sont les **coefficients de régression**. ϵ est une variable aléatoire représentant un résidu (erreur de mesure). En effet, ce modèle suppose que la variable expliquée que l'on observée résulte du vrai modèle (ici le modèle linéaire représentée par la droite) et un bruit (de mesure par exemple) ou tout autre type d'erreur. Ce bruit est généralement supposé d'espérance nulle et de

1. En informatique et en particulier en machine learning (apprentissage), on trouve aussi l'appellation entrées/sorties pour respectivement variables explicatives et expliquées.

variance σ^2 et décorrélé ($\text{cov}(\epsilon_i, \epsilon_j)_{i \neq j} = 0$). Dans ce cas σ^2 devient également un paramètre du modèle et est donc aussi à estimer. Dans le cadre de ce cours, on va supposer que ce bruit est en plus Gaussien. Il en découle donc qu'il est indépendant (les ϵ_i sont i.i.d).

Les deux paramètres (β_0, β_1) sont inconnus et donc à estimer. Cette estimation sera effectuée à partir d'un échantillon de couples $((x_1, Y_1), \dots, (x_n, Y_n))^2$. Le modèle s'écrit donc sous la forme

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad (8.2)$$

où ϵ_i est le bruit associée à la i ème variable aléatoire. Étant donné donc une réalisation (valeur) y_i de chaque Y_i et le résidu associé (réalisation de la variable aléatoire représentant le bruit) que nous notons e_i on obtient alors :

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad (8.3)$$

Le modèle de régression linéaire est souvent estimé par la méthode des moindres carrés mais il existe aussi de nombreuses autres méthodes pour estimer ce modèle. On peut par exemple estimer le modèle par maximum de vraisemblance ou encore par inférence bayésienne (maximum a posteriori).

8.2.1 Estimation par moindres carrés

La méthode des moindres carrés est une approche d'estimation ponctuelle des paramètres de régression (β_0, β_1) . Elle consiste à fournir les estimations $(\hat{\beta}_0, \hat{\beta}_1)$ qui minimisent la somme des écarts quadratiques (somme des carrés résidus) entre les valeurs observées y_i et l'espérance $\beta_0 + \beta_1 x_i$ du modèle de Y_i . D'après l'équation (8.3), l'écart entre la valeur d'une observation et l'espérance du modèle est donné par

$$e_i = y_i - (\beta_0 + \beta_1 x_i).$$

La somme des carrés des résidus est donc donnée par

$$\text{RSS}(\beta_0, \beta_1) = Q = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2. \quad (8.4)$$

La fonction de deux variables Q est une fonction quadratique et sa minimisation, comme nous allons le voir ci-dessous, peut s'effectuer facilement de façon analytique.

Théorème 8.2.1. *Considérons le modèle de régression simple donné par l'équation (8.1). Soit $((x_1, y_1), \dots, (x_n, y_n))$ un échantillon de valeurs observées de Y et leurs valeurs associées de x . Alors les estimations des moindres carrés ordinaires (MCO) de β_0 et β_1 sont données par :*

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad (8.5)$$

$$\hat{\beta}_1 = \left[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \right] \left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^{-1}, \quad (8.6)$$

2. Ici nous utilisons la notations (x_i, Y_i) vu que x est déterministe mais cela ne change rien au modèle si X est aléatoire.

où \bar{x} représente la moyenne empirique des x_i et \bar{y} la moyenne empirique des y_i et sont donnés par

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

et

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

Preuve. Selon les MCO, les estimations $\hat{\beta}_0$ et $\hat{\beta}_1$ sont celles qui minimisent la somme des carrés des résidus (8.4) par rapport à $\hat{\beta}_0$ et $\hat{\beta}_1$. Mathématiquement on note

$$\begin{aligned} (\hat{\beta}_0, \hat{\beta}_1) &= \arg \min_{(\beta_0, \beta_1) \in \mathbb{R}^2} Q(\beta_0, \beta_1) \\ &= \arg \min_{(\beta_0, \beta_1) \in \mathbb{R}^2} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2. \end{aligned} \quad (8.7)$$

Minimiser la fonction RSS revient à prendre les paramètres (β_0, β_1) qui annulent sa dérivée première et pour lesquels la dérivée second est positive. Nous aurons donc déterminé les estimations $(\hat{\beta}_0, \hat{\beta}_1)$. Ainsi, nous avons :

$$\begin{aligned} \frac{\partial Q}{\partial \beta_0} &= \frac{\partial \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i)) \\ \frac{\partial Q}{\partial \beta_1} &= \frac{\partial \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2}{\partial \beta_1} = -2 \sum_{i=1}^n x_i (y_i - (\beta_0 + \beta_1 x_i)). \end{aligned}$$

Ensuite, en annulant ces dérivés nous avons :

$$\begin{aligned} \sum_{i=1}^n y_i - \sum_{i=1}^n \hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^n x_i &= 0 \\ \sum_{i=1}^n x_i y_i - \hat{\beta}_0 \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 &= 0. \end{aligned}$$

ce qui donne

$$\begin{aligned} n\hat{\beta}_0 &= \sum_{i=1}^n y_i - \hat{\beta}_1 \sum_{i=1}^n x_i \\ \hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n x_i y_i \end{aligned}$$

qu'on appelle les *équations normales*. La première équation donne

$$n\hat{\beta}_0 = \sum_{i=1}^n y_i - \hat{\beta}_1 \sum_{i=1}^n x_i$$

et en divisant par n on obtient la valeur de $\hat{\beta}_0$

$$\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^n y_i - \hat{\beta}_1 \left(\frac{1}{n} \sum_{i=1}^n x_i \right) = \bar{y} - \hat{\beta}_1 \bar{x}.$$

La seconde équation donne

$$\hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i.$$

En remplaçant $\hat{\beta}_0$ par sa valeur on obtient

$$\sum_{i=1}^n x_i \bar{y} - \hat{\beta}_1 \sum_{i=1}^n x_i \bar{x} + \hat{\beta}_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i,$$

ce qui donne enfin

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \bar{y}}{\sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \bar{x}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Maintenant il faut vérifier si la dérivée partielle seconde de Q par rapport à au moins l'un des paramètres est positive (minimum de la fonction minimisée) et que le déterminant de la matrice des dérivées partielles secondes de Q est strictement positif. Partons de la dérivée partielle dans (8.8), on obtient

$$\begin{aligned} \frac{\partial^2 Q}{\partial^2 \beta_0} &= \frac{-2 \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))}{\partial \beta_0} \\ &= -2 \sum_{i=1}^n \frac{\partial \beta_0}{\partial \beta_0} = 2n \geq 0. \end{aligned} \quad (8.8)$$

Maintenant il reste à vérifier que le déterminant de la matrice des dérivées partielles secondes de Q est strictement positif. Le déterminant de cette matrice est donné par

$$\det \begin{pmatrix} \frac{\partial^2 Q}{\partial^2 \beta_0} & \frac{\partial^2 Q}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 Q}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 Q}{\partial^2 \beta_1} \end{pmatrix} = \det \begin{pmatrix} 2n & 2 \sum_i x_i \\ 2 \sum_i x_i & 2 \sum_i x_i^2 \end{pmatrix} = 4n \sum_i (x_i - \bar{x})^2 > 0.$$

Notez que ce déterminant est nul si tous les x_i prennent la même valeur. Ainsi, au moins deux valeurs x_i distinctes sont nécessaires pour la détermination des coefficients de régressions (β_0, β_1) .

8.2.2 Formulation vectorielle

maintenant nous reformulons le modèle que nous venons de voir sous forme de vecteurs-matrices. Comme nous allons le voir, les résultats sous la forme matricielle sont obtenus à partir de calculs simples. Cela permettra aussi de généraliser le modèle linéaire simple à des modèles généraux notamment la régression multiple.

Soit (y_1, \dots, y_n) l'ensemble des valeurs observées de la variable dépendante Y et (x_1, \dots, x_n) l'ensemble des valeurs observées de la variable explicative x .

▲ Remarque 8.2.1. L'ensemble des couples $((x_1, y_1), \dots, (x_n, y_n))$ s'appelle aussi ensemble d'apprentissage. Car c'est l'ensemble de données à partir duquel on va estimer notre modèle (donc apprendre le modèle) pour pouvoir ensuite prédire la valeur de Y pour une nouvelle valeur de x

Selon le modèle de régression linéaire simple on a

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1, \\ y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2, \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_n + \epsilon_n. \end{aligned} \quad (8.9)$$

Si l'on note par $\mathbf{y} = (y_1, \dots, y_n)^T$ le vecteur des valeurs d'observations de Y , $\mathbf{e} = (\epsilon_1, \dots, \epsilon_n)^T$ le vecteur des valeurs des résidus, $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ le vecteur paramètre à estimer, et enfin par \mathbf{X} la *matrice de régression* (appelée aussi matrice de *design* ou de *Vandermonde*)

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad (8.10)$$

le modèle (8.9) s'écrit donc sous la forme matricielle suivante :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}. \quad (8.11)$$

La somme des écarts quadratiques donnée par l'équation (8.4) est maintenant donnée par :

$$\text{RSS}(\boldsymbol{\beta}) = Q(\boldsymbol{\beta}) = \mathbf{e}^T \mathbf{e} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (8.12)$$

$$= \mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X}\boldsymbol{\beta}. \quad (8.13)$$

L'estimation $\hat{\boldsymbol{\beta}}$ par moindres carrés de $\boldsymbol{\beta}$ s'obtient en minimisant (8.13) qui est une fonction quadratique en $\boldsymbol{\beta}$. En dérivons (8.13) par rapport à $\boldsymbol{\beta}$ et en annulant cette dérivée on obtient

$$-2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X}\boldsymbol{\beta} = \mathbf{0} \quad (8.14)$$

et donc les équations normales

$$\mathbf{X}^T \mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T \mathbf{y}. \quad (8.15)$$

En multipliant cette équation par $(\mathbf{X}^T \mathbf{X})^{-1}$ on obtient l'estimation de $\boldsymbol{\beta}$

$$\boldsymbol{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \quad (8.16)$$

Notez que l'inverse $(\mathbf{X}^T \mathbf{X})^{-1}$ existe si les données comportent au moins deux valeurs distinctes de x_i .

Chapitre 9

Estimation par intervalle

Nous allons maintenant voir une autre approche d'estimation des paramètres *l'estimation par intervalle*. L'estimation par intervalle fournit, à partir d'un échantillon d'une population, non seulement des valeurs des paramètres à estimer, mais un intervalle de valeurs centré sur la valeur numérique estimée du paramètre inconnu avec un *niveau de confiance* donné. Ce niveau de confiance représente la probabilité que le vrai paramètre se trouve dans l'intervalle que l'on donne comme estimation. L'intervalle est appelé *intervalle de confiance* le niveau de confiance est aussi appelé *précision* ou *coefficient de confiance*.

9.1 Intervalle de confiance

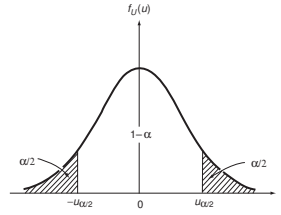
Définition 9.1.1. Intervalle de confiance. Soit (X_1, \dots, X_n) un échantillon de variables aléatoires issues d'une population de densité $f(x; \theta)$, θ étant le (vecteur) paramètre à estimer. Supposons aussi que $T_1(X_1, \dots, X_n)$ et $T_2(X_1, \dots, X_n)$ deux statistiques sur l'échantillon telle que $T_1 < T_2$. L'intervalle $[T_1, T_2]$ est dit *intervalle de confiance* à $100(1 - \alpha)\%$ pour θ si

$$P[T_1 < \theta < T_2] = 1 - \alpha. \quad (9.1)$$

α représente le *risque* que le vrai paramètre θ ne soit pas dans cet intervalle et $1 - \alpha$ s'appelle *niveau* ou (*coefficient*) *de confiance*. Une estimation par intervalle de confiance sera donc d'autant meilleure que l'intervalle sera petit pour un coefficient de confiance grand (proche de 1) ou de manière équivalente pour un risque α proche de zéro. Les valeurs généralement prises pour $1 - \alpha$ sont 0.90, 0.95, 0.99, and 0.999. Les limites de l'intervalle T_1 et T_2 sont appelés respectivement la *limite inférieure de confiance* et *limite supérieure de confiance*.

△ Remarques

- L'intervalle de confiance est fonction de l'estimation du paramètre θ
- L'intervalle de confiance est également fonction de α . A taille d'échantillon n fixée, lorsqu'on augmente le niveau de confiance $1 - \alpha$, la largeur de l'Intervalle de Confiance (IC)
- Pour un niveau de confiance $1 - \alpha$ fixé, lorsqu'on augmente la taille de l'échantillon n , la largeur de l'IC diminue.

FIGURE 9.1 – Loi normale centrée réduite et IC à $[100(1 - \alpha)]\%$

9.1.1 Calcul d'un intervalle de confiance

Soit a et b les bornes d'un intervalle de confiance $IC_{1-\alpha}(\theta)$ pour le paramètre θ on a On a :

$$\mathbb{P}(a < \theta < b) = 1 - \alpha \implies \mathbb{P}(\theta < a) + \mathbb{P}(\theta > b) = \alpha \quad (9.2)$$

En posant $\alpha = \alpha_1 + \alpha_2$, il existe donc une infinité de choix possibles pour α_1 et α_2 , et donc de choix pour a et b et donc de l'IC. Pour l'instant, nous ne considérons que le cas d'un intervalle de confiance bilatéral symétrique, où on a les mêmes risques $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$

Notons que, connaissant la loi de l'estimateur, il est possible de donner un intervalle de confiance. Ici nous considérons les intervalles de confiance les plus classiques.

9.1.2 Loi normale : Intervalle de confiance sur μ

Loi normale : Intervalle de confiance pour μ avec σ connu

On a vu que \bar{X} est le meilleur estimateur de μ et que

$$U = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim \mathcal{N}(0, 1) \quad (9.3)$$

En prenant des risques symétriques ($\frac{\alpha}{2}$), pour un risque fixé, on peut donc lire dans les tables de probabilités de la loi normale centrée réduite, le **quantile** $\mathbb{P}(U \leq \frac{\alpha}{2})$

Remarque : Comme le risque est symétrique ici, on a donc

$$\mathbb{P}(U \geq \frac{\alpha}{2}) = \alpha - \mathbb{P}(U \leq \frac{\alpha}{2}) = \frac{\alpha}{2} \quad (9.4)$$

La notion de quantile est définie de la façon suivante :

Définition 9.1.2. pour une variable aléatoire continue X , le quantile q_α d'ordre α de la loi de X est telle que

$$\mathbb{P}(U \leq q_\alpha) = \alpha \quad (9.5)$$

Remarque : la notation généralement utilisée pour les quantile est : u_α pour la loi normale, t_α pour la loi de Student à n degrés de liberté, χ_α^n pour la loi χ_n^2 , etc

Le risque étant symétrique, d'après (9.4) on a

$$\mathbb{P}(-u_{\frac{\alpha}{2}} \leq U \leq u_{\frac{\alpha}{2}}) = 1 - \alpha \quad (9.6)$$

et d'après (9.3) on obtient

$$\mathbb{P}\left(\bar{X} - u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \quad (9.7)$$

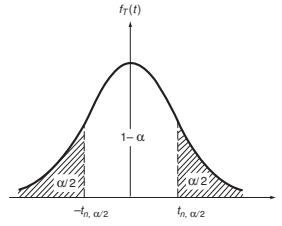


FIGURE 9.2 – Loi de Student à $n - 1$ degrés de liberté et IC à $[100(1 - \alpha)]\%$

d'où l'intervalle de confiance sur μ :

$$IC_{1-\alpha}(\mu) = \left[\bar{X} - u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right] \quad (9.8)$$

exemple pour $\alpha = 0.05$

$$IC_{0.95}(\mu) = \left[\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right] \quad (9.9)$$

Loi normale : Intervalle de confiance pour μ avec σ inconnu

Pour calculer l'IC sur μ on a vu que la statistique à utiliser est $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$. Or, comme la variance σ^2 est inconnue, on utilise à sa place son meilleur estimateur : la variance empirique corrigée $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_i)^2$

La statistique à utiliser est donc

$$T = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \quad (9.10)$$

On sait que $Z = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$

la statistique (9.10) pour calculer l'IC s'écrit dont

$$T = \frac{U}{\sqrt{\frac{Z}{n-1}}} \sim \text{Loi de Student à } n-1 \text{ degrés de liberté} \quad (9.11)$$

Soit $t_{n-1, \frac{\alpha}{2}} = \mathbb{P}(T \leq \frac{\alpha}{2})$ le quantile d'ordre $\alpha/2$ de la loi de Student à $n - 1$ degrés de liberté.

L'intervalle de confiance est donc donné par

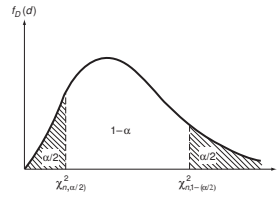
$$\mathbb{P}(-t_{n-1, \frac{\alpha}{2}} \leq T \leq t_{n-1, \frac{\alpha}{2}}) = 1 - \alpha \quad (9.12)$$

On obtient donc l'intervalle de confiance comme précédemment

$$IC_{1-\alpha}(\mu) = \left[\bar{X} - t_{n-1, \frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1, \frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right] \quad (9.13)$$

où $t_{n-1, \frac{\alpha}{2}}$ est le quantile d'ordre $\alpha/2$ de la loi de Student à $n - 1$ degrés de liberté

⚠ Remarque 9.1.1. Si la loi de X n'est pas normale, on sait d'après le théorème central limite que lorsque la taille d'échantillon est grande, \bar{X} suit une loi normale, et donc les résultats précédents sont applicables.

FIGURE 9.3 – Loi de χ^2 à n degrés de liberté et IC à $[100(1 - \alpha)]\%$

9.1.3 Loi normale : Intervalle de confiance sur σ^2

Loi normale : Intervalle de confiance pour σ lorsque μ connu

On sait que la variance observée $V^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$ constitue le meilleur estimateur de σ^2 lorsque μ est connue.

D'autre part :

$$D = \frac{n}{\sigma^2} V^2 \sim \chi_n^2 \quad (9.14)$$

Soit $\chi_{n, \frac{\alpha}{2}}^2 = \mathbb{P}(D \leq \frac{\alpha}{2})$ le quantile d'ordre $\frac{\alpha}{2}$ de la loi de χ^2 à n degrés de liberté.

l'IC $_{1-\alpha}(\sigma^2)$ est donc donné par

$$\mathbb{P}\left(\chi_{n, \frac{\alpha}{2}}^2 \leq D = \frac{n}{\sigma^2} V^2 \leq \chi_{n, 1-\frac{\alpha}{2}}^2\right) = 1 - \alpha$$

donc

$$\mathbb{P}\left(\frac{nV^2}{\chi_{n, 1-\frac{\alpha}{2}}^2} \leq \sigma^2 \leq \frac{nV^2}{\chi_{n, \frac{\alpha}{2}}^2}\right) = 1 - \alpha$$

finalement on obtient

$$\text{IC}_{1-\alpha}(\sigma^2) = \left[\frac{nV^2}{\chi_{n, 1-\frac{\alpha}{2}}^2}, \leq \frac{nV^2}{\chi_{n, \frac{\alpha}{2}}^2} \right] \quad (9.15)$$

Loi normale : Intervalle de confiance pour σ lorsque μ inconnu

Lorsque μ est inconnue, on sait que que la variance empirique corrigée $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ constitue le meilleur estimateur de σ^2

On sait également que :

$$D = \frac{n-1}{\sigma^2} S^2 \sim \chi_{n-1}^2 \quad (9.16)$$

l'IC $_{1-\alpha}(\sigma^2)$ est donc donné par

$$\mathbb{P}\left(\chi_{n-1, \frac{\alpha}{2}}^2 \leq D = \frac{n-1}{\sigma^2} S^2 \leq \chi_{n-1, 1-\frac{\alpha}{2}}^2\right) = 1 - \alpha$$

donc

$$\text{IC}_{1-\alpha}(\sigma^2) = \left[\frac{(n-1)S^2}{\chi_{n-1, 1-\frac{\alpha}{2}}^2}, \frac{(n-1)S^2}{\chi_{n-1, \frac{\alpha}{2}}^2} \right] \quad (9.17)$$

Remarque : Ces intervalles de confiance sur la variance ne sont valables que pour une loi normale. Contrairement au cas de la moyenne, ces résultats ne peuvent être étendus aux cas d'autres lois

Chapitre 10

Tests d'hypothèses

Dans cette partie nous allons étudier le *test statistique d'hypothèse*. Un *test statistique* est un procédé qui permet de décider entre deux ou plusieurs hypothèses sur une population selon les résultats obtenus à partir d'un échantillon de cette population. Généralement, on teste une hypothèse sur laquelle on se demande si les données observées fournissent une évidence suffisante pour la rejeter, sinon elle est retenue. Cette hypothèse s'appelle l'*hypothèse nulle* et se note H_0 . Par exemple, si le test concerne la valeur d'un paramètre θ , cette hypothèse nulle peut s'écrire

$$H_0 : \theta \in \Theta_0 \quad (10.1)$$

où Θ_0 est l'ensemble de valeurs supposée du paramètre θ selon H_0 . Toute autre hypothèse qui diffère de l'hypothèse nulle s'appelle l'*hypothèse alternative* (ou contre-hypothèse) qui se note H_1 . L'hypothèse nulle est donc testée contre l'hypothèse alternative.

Une hypothèse est dite *simple* si elle ne contient qu'un seul élément, ce qui est généralement le cas pour $H_0 : \theta = \theta_0$; sinon elle est composite. L'hypothèse alternative est généralement composite

$$H_1 : \theta \in \Theta_1 \quad (10.2)$$

avec Θ_1 un sous ensemble de l'ensemble des paramètres disjoint de θ_0 . H_1 se ramène souvent aux trois cas suivants

1. $H_1 : \theta < \theta_0$,
2. $H_1 : \theta > \theta_0$,
3. $H_1 : \theta \neq \theta_0$.

Dans les deux premiers cas, le test est dit *unilatéral* et dans le dernier le test est dit *bilatéral*.

10.1 Région de rejet d'un test

Soit X une variable aléatoire et \mathcal{X} l'ensemble de ses valeurs. Le test s'effectue en trouvant un sous-ensemble $R \subseteq \mathcal{X}$ appelé la *région de rejet*. Ainsi, si $X \in R$,

l'hypothèse nulle (H_0) est rejetée, sinon, elle est retenue. Cette région se définit sous la forme suivante

$$R = \{x : T(x) > s\} \quad (10.3)$$

où T est une *statistique de test* et s un *seuil*. le problème en test d'hypothèse est donc de trouver une statistique de test convenable et une valeur convenable pour le seuil de rejet s .

10.2 Erreurs associées à un test

Bien sur, en effectuant un test d'hypothèses, on peut se tromper en rejetant l'hypothèse nulle ou en l'acceptant. Il existe donc deux types d'erreur : l'erreur de première espèce (dite aussi erreur de type I) et l'*erreur de deuxième espèce* (dite aussi erreur de type II). L'erreur de première espèce correspond au cas où l'on rejette H_0 (décider H_1) alors que celle-ci est vraie. L'erreur de deuxième espèce correspond quant à elle au cas où l'on rejette H_1 (décider H_0) alors que celle-ci est vraie. Les décisions possibles sont récapitulées par le tableau suivant. Pour chacune des deux erreurs, on associe un une

Décision \ Vérité	H_0	H_1
H_0	décision correcte	erreur de deuxième espèce
H_1	erreur de première espèce	décision correcte

TABLE 10.1 – Récapitulatif des décisions en test d'hypothèse

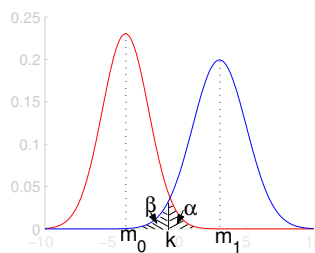
probabilité (*un risque*). Le risque de première espèce est noté α . Il représente le risque de rejeter H_0 à tort. Des valeurs de ce risque sont 1%, 5%, 10% qui correspondent aux niveaux de confiance 99%, 95% et 90%. Le *niveau de confiance du test* est donc $1 - \alpha$ qui correspond à retenir H_0 à raison. Le risque de deuxième espèce représente quant à lui le risque de retenir H_0 à tort. Il est noté β .

La *puissance d'un test* est la probabilité de rejeter l'hypothèse nulle à raison. La puissance du test est donc le complément de l'erreur de deuxième espèce et est donc égale à $1 - \beta$.

On peut résumer cela par le tableau suivant :

Décision \ Vérité	H_0	H_1
H_0	niveau de confiance $1 - \alpha$	risque β
H_1	risque α	puissance de test $1 - \beta$

TABLE 10.2 – Récapitulatif sur les risques associés à un test d'hypothèses



10.3 Statistiques de test

Le choix de la statistique de test et de la région de rejet s'effectue de façon à maximiser la puissance du test $1 - \beta$ pour un risque de première espèce α fixé.

10.3.1 Test du rapport de vraisemblance

Si l'on se place dans le cadre d'un test entre deux hypothèses simples

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1,$$

Le théorème de Neyman et Pearson montre que le *test du rapport de vraisemblance* est le test le plus puissant avec un risque α . Selon ce test, la région critique (de rejet) optimale est définie par

$$R = \left\{ x : \frac{L(\theta_1; x)}{L(\theta_0; x)} > s_\alpha \right\}$$

avec $L(\theta_k; x)$ étant la vraisemblance de θ_k pour x . Le seuil de rejet s_α , qui dépende de α , est déterminé par $\alpha = \mathbb{P}_{\theta_0}(X \in R)$.

Test du rapport de vraisemblance (ou Test de Neyman-Pearson) : Exemple

Soit un échantillon d'observations *i.i.d.* (x_1, \dots, x_n) où $X_i \sim \mathcal{N}(x_i; \mu, \sigma^2)$ supposons que la variance σ^2 est connue et que l'espérance μ est inconnue. Considérons le test

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu = \mu_1,$$

avec $\mu < \mu_1$

On la vraisemblance de μ pour l'échantillon $\mathbf{x} = (x_1, \dots, x_n)$ est

$$\begin{aligned} L(\mu; \mathbf{x}) &= \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2} \left(\frac{x_i - \mu}{\sigma}\right)^2\right] \\ &= \frac{1}{(\sigma^2\sqrt{2\pi})^n} \exp\left[-\sum_{i=1}^n \frac{1}{2\sigma^2} (x_i - \mu)^2\right] \end{aligned} \quad (10.4)$$

le rapport de vraisemblance est donc donné par

$$\frac{L(\mu_1; \mathbf{x})}{L(\mu_0; \mathbf{x})} = \exp\left[\frac{1}{2\sigma^2} 2(\mu_1 - \mu_0) \sum_{i=1}^n x_i - \frac{n}{2\sigma^2} (\mu_1^2 - \mu_0^2)\right] \quad (10.5)$$

Donc, en prenant le logarithme, $\frac{L(\mu_1; \mathbf{x})}{L(\mu_0; \mathbf{x})} > s_\alpha$ est équivalent à

$$\bar{x} > \log(s_\alpha) \frac{\sigma^2}{n(\mu_1 - \mu_0)} + \frac{(\mu_1 + \mu_0)}{2} = \text{cste}$$

On a vu que cette *cste* qui dépende α est déterminé par $\alpha = \mathbb{P}_{\mu_0}(X \in R)$ qui vaut dans ce cas $\alpha = \mathbb{P}_{\mu_0}(\bar{x} > \text{cste})$

La région de rejet du test est donc donnée par

$$R = \left\{ \mathbf{x} : \bar{x} > \mu_0 + u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right\} \quad (10.6)$$

10.3.2 Test de Wald

Soit θ un paramètre et soit $\hat{\Theta}$ un estimateur de ce paramètre et soit $\hat{\sigma}$ l'écart type de cet estimateur $\hat{\Theta}$. Considérons le test

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0,$$

Supposons que $\hat{\Theta}$ est asymptotiquement normal : $\frac{\sqrt{n}(\hat{\Theta} - \theta_0)}{\sqrt{\text{var}(\hat{\Theta})}} \xrightarrow[n \rightarrow \infty]{\text{loi}} \mathcal{N}(0, 1)$. Dans le *test de Wald*, la statistique de test est donnée par

$$\frac{\hat{\Theta} - \theta_0}{\sqrt{\text{var}(\hat{\Theta})}} \quad (10.7)$$

où $\sqrt{\text{var}(\hat{\Theta})}$ représente l'écart type de l'estimateur. Le test de Wald consiste à comparer cette statistique à la loi normale centrée réduite. Il consiste alors à rejeter H_0 si

$$\left| \frac{(\hat{\Theta} - \theta_0)}{\sqrt{\text{var}(\hat{\Theta})}} \right| > u_{\alpha/2}$$

où $u_{\alpha/2}$ est le quantile d'ordre $\alpha/2$ de la loi normale centrée réduite.

Exemple : cas d'un grand échantillon gaussien

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0,$$

lorsque σ est connue

la statistique de test sous H_0 dans ce cas est donnée par

$$U = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \quad (10.8)$$

On sait que $U \sim \mathcal{N}(0, 1)$

On rejette donc H_0 si

$$\left| \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \right| > u_{\frac{\alpha}{2}}$$

où $u_{\frac{\alpha}{2}}$ est le quantile d'ordre $\alpha/2$ de la loi normale centrée réduite
ou par équivalence : rejeter H_0 si

$$|\bar{X} - \mu_0| > u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$