L3 Miashs/Maths
Statistiques et probabilités 3

Faicel Chamroukhi
Professeur



email: chamroukhi@unicaen.fr
web: chamroukhi.com

2018

Plan I

Notions 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 : 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.

Cadre de la statistique :

Definition

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

Statistique descriptive :

Definition

L'objectif de la statistique descriptive est de décrire, résumer ou représenter, par des statistiques, les données disponibles dans un échantillon

Wikipedia : Descriptive statistics are distinguished from inferential statistics (or inductive statistics), in that descriptive statistics aim to summarize a sample, rather than use the data to learn about the population that the sample of data is thought to represent.

Expérience aléatoire

On appelle **expérience aléatoire** une expérience dont on connaît l'ensemble des résultats possibles mais dont on ne peut prédire le résultat effectif avec certitude et 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*. On appelle *événement* un 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 $\mathbb{P} : \mathcal{A} \to [0, 1]$, appelée mesure ou distribution de probabilité, qui à tout événement A associe un nombre $\mathbb{P}(A)$ appelé probabilité de cet événement.

Variables aléatoires

- Variables aléatoires
- Couple de variables aléatoires
- Vecteurs Aléatoires
- Variables et vecteurs aléatoires gaussiens

Variables aléatoires

Une **Variable statistique** X est une application de Ω dans un ensemble E
$X : \Omega \longrightarrow E$

Types de variables aléatoires

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 representée par une qualité, 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**.

Variables aléatoires

Definition

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

Exemple : le résultat du jet d'un dé, le nombre d'enfants dans une famille, sont des variables aléatoires discrètes.

Definition

Une variable aléatoire est dite **continue** 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.

Variables aléatoires

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

Definition

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

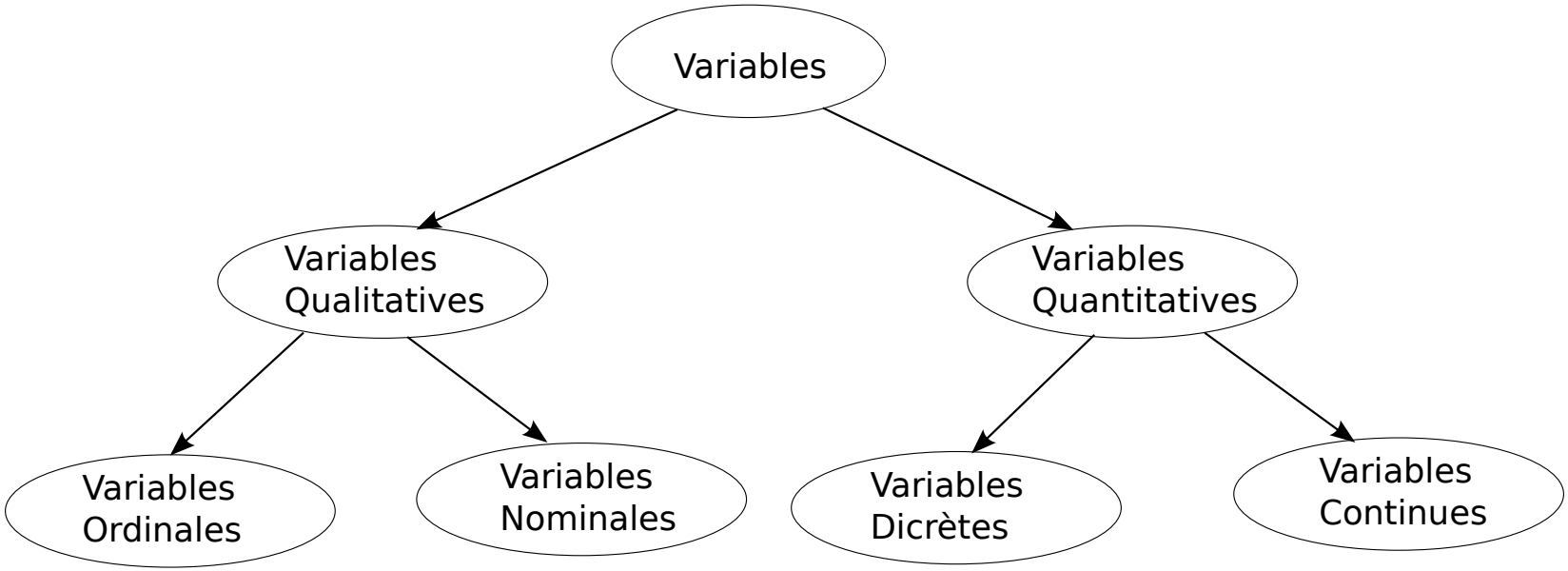
Par exemple, la variable sexe est une variable qualitative nominale qui a deux modalités possibles : féminin ou masculin et dont l'ordre n'importe pas. Un autre exemple est la catégorie socioprofessionnelle

Definition

Une variable aléatoire qualitative **ordinale** est une variable qui correspondent à un ordre. Il y a un ordre sur les modalités.

Par exemple, le degré de satisfaction par rapport à un service : très satisfait, satisfait, insatisfait, etc

Variables aléatoires : récapitulatif



Fonction de densité de probabilité

Definition

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

$$\begin{array}{rcl} f : \mathbb{R} & \rightarrow & \mathbb{R}^+ \\ x & \rightarrow & f(x) \end{array} \quad (1)$$

telle que :

$$(1) \qquad \forall x \in \mathbb{R} \quad f(x) \geq 0$$

$$(2) \qquad \int_{\mathbb{R}} f(x)dx = 1$$

Fonction de répartition

La fonction de répartition d'une variable aléatoire réelle X, notée $F_X(x)$, caractérise la loi de probabilité de cette variable. Elle représente la probabilité que la variable aléatoire réelle X prenne une valeur inférieure ou égale à x :

$$F_X(x) = \mathbb{P}(X \leq x) \tag{2}$$

et, pour tout nombre réel x, est donnée par :

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,\mathrm{d}u. \tag{3}$$

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

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

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

Espérance mathématique d'une v.a

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

Espérance mathématique d'une variable aléatoire discrète

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

$$\mathbb{E}[X] = \sum_k x_k P(X = x_k) = \sum_k x_k p_k \tag{4}$$

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

Espérance mathématique d'une v.a

Espérance mathématique d'une variable aléatoire discrète

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

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f_X(x) dx. \tag{5}$$

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

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x) f_X(x) dx \tag{6}$$

Variance d'une variable aléatoire

Definition

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

$$var[X] = \mathbb{E}[(X - E[X])^2] = \mathbb{E}[X^2] - (E[X])^2 \tag{7}$$

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 : D'après (6) et (7), la variance est donc aussi donnée par

$$var[X] = \int_{\mathbb{R}} (x - E[X])^2 f_X(x) dx. \tag{8}$$

Quelques propriétés de l'espérance et de la variance

Soit X une v.a réelle, deux fonctions $g, h : \mathbb{R} \to \mathbb{R}$ 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 :

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

$\Rightarrow$ Propriété de **linéarité** de l'espérance

Soient n variables aléatoires réelles $X_1, \ldots, X_n$. Alors $\mathbb{E}[\sum_i X_i] = \sum_i \mathbb{E}[X_i]$

Quelques propriétés de l'espérance et de la variance

De même, pour la variance, on a la propriété suivante :

$$Var(aX + b) = a^2 Var(X)$$

$\Rightarrow$ Propriété **quadratique** de la variance

⚠ Remarque : 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

Definition

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

moyenne empirique

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{9}$$

variance empirique

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_i)^2 . \tag{10}$$

Quelques statistique sur variables aléatoires

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 :

$$M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \,. \tag{11}$$

Couple de variables aléatoires

- Variables aléatoires
- Couple de variables aléatoires
- Vecteurs Aléatoires
- Variables et vecteurs aléatoires gaussiens

Couple de variables aléatoires, lois jointes, loi conditionnelles, loi marginales, notion d'indépendance I

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. $\Rightarrow$ Cette information est représentée par la distribution de Y conditionnellement à $X = x$ (distribution de Y étant donné à $X = x$) soit $f_{Y|X}(y|x)$.

Théorème de Bayes

Ceci est explicité par le théorème de suivant, appelé **théorème de Bayes** (ou aussi règle de Bayes ou formule de Bayes) comme suit

Theorem

$$f_{XY}(x,y) = f_{Y|X}(y|x) f_X(x) \tag{12}$$

et par symétrie on a également

$$f_{XY}(x,y) = f_{X|Y}(x|y) f_Y(y)\ . \tag{13}$$

Pour les VA discrètes cela revient à

Theorem

$$\begin{aligned} P(Y=y, X=x) &= P(Y=y|X=x)P(X=x) && (14)\\ &= P(X=x|Y=y)P(Y=y) && (15) \end{aligned}$$

Loi jointe de couple 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.

Definition

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 : 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	$\ldots$	x	$\ldots$	Somme des colonnes
$\vdots$				
y		$P(X = x, Y = y)$		$P(Y = y)$
$\vdots$				
Somme des lignes		$P(X = x)$		

Loi marginale de v.a discrètes

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

$$P(X = x) = \sum_{y \in B} P(X = x, Y = y) \tag{16}$$

et

$$P(Y = y) = \sum_{x \in A} P(X = x, Y = y) \tag{17}$$

⚠ Remarque : On peut remarquer que ces lois correspondent respectivement à la somme des lignes et la somme des colonnes dans le tableau précédent (1).

Loi jointe de couple de v.a continues (à densité)

Definition

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

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u, v) du dv \tag{18}$$

satisfaisant les conditions suivantes :

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

Densités marginales

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

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)dy \tag{19}$$

et

$$f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x,y)dx. \tag{20}$$

Thm de changement de variables

(théorème du changement de variable) Soient (X, Y) un couple de *var* de densité $f_{(X,Y)}$, et $\phi : (X, Y)(\Omega) \to \mathbb{R}$ et $\psi : (X, Y)(\Omega) \to \mathbb{R}$ deux fonctions. On considère le couple de *var* :

$$(U, V) = (\phi(X, Y), \psi(X, Y)).$$

On suppose que $g(x, y) = (\phi(x, y), \psi(x, y))$ est injective et qu'il existe deux fonctions $h : \mathbb{R}^2 \to X(\Omega)$ et $k : \mathbb{R}^2 \to Y(\Omega)$ différentiables telles que

$$\begin{cases} u = \phi(x, y), \\ v = \psi(x, y), \end{cases} \quad \Leftrightarrow \quad \begin{cases} x = h(u, v), \\ y = k(u, v). \end{cases}$$

Soit $J(u, v)$ le jacobien associé à $(x, y) = (h(u, v), k(u, v))$ défini par le déterminant :

$$J(u, v) = \begin{vmatrix} \frac{\partial h}{\partial u}(u, v) & \frac{\partial h}{\partial v}(u, v) \\ \\ \frac{\partial k}{\partial u}(u, v) & \frac{\partial k}{\partial v}(u, v) \end{vmatrix} = \frac{\partial h}{\partial u}(u, v)\frac{\partial k}{\partial v}(u, v) - \frac{\partial k}{\partial u}(u, v)\frac{\partial h}{\partial v}(u, v).$$

Alors une densité de (U, V) est donnée par

$$f_{(U,V)}(u, v) = f_{(X,Y)}(h(u, v), k(u, v))|J(u, v)|, \qquad (u, v) \in \mathbb{R}^2.$$

Covariance de deux variables aléatoires

Pour une variable aléatoire, on parle de variance

pour un couple de deux variables aléatoires, on parle de **covariance**

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

Definition

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

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \tag{21}$$

et on la propriété suivante

Definition

$$cov(X, Y) = \sigma_{XY} = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \tag{22}$$

Covariance de deux variables aléatoires I

⚠ Remarque : 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

$$\mathbb{E}[XY] = \sum_{x \in A} \sum_{y \in B} xyP(X = x, Y = y) \tag{23}$$

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

$$\mathbb{E}[XY] = \int_{\mathbb{R}} \int_{\mathbb{R}} xyf_{X,Y}(x, y)\, \mathrm{d}\, x\, \mathrm{d}\, y. \tag{24}$$

Corrélation entre deux variables aléatoires I

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

Definition

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

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}. \tag{25}$$

On a toujours $\rho_{X,Y} \in [-1,1]$. Plus $|\rho_{X,Y}|$ est proche de 1 plus la corrélation entre les variables X et Y est forte. *Si X et Y sont indépendantes, alors $Cov(X,Y) = 0$ et donc $\rho_{X,Y} = 0$. On a par conséquent, $Var(X+Y) = Var(X) + Var(Y)$.*

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).

Definition

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

$$f_{XY}(x, y) = f_X(x) f_Y(y) \tag{26}$$

Pour les VA discrètes cela revient à

$$P(Y = y, X = x) = P(X = x) P(Y = y) \tag{27}$$

Indépendance de deux variables aléatoires

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) $f_{XY}(x, y)$

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 $Y \Rightarrow$ 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.

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

$$f_{XY}(x, y) = f_{Y|X}(y|x) f_X(x) \tag{28}$$

par le théorème de Bayes.

Indépendance de deux variables aléatoires

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

$$f_Y(y) = \int_{\mathcal{X}} f_{XY}(x,y)dx = \int_{\mathcal{X}} f_{Y|X}(y|x)f_X(x) \tag{29}$$

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

$$f_Y(y) = f_{Y|X}(y|x) \int_{\mathcal{X}} f_X(x) = f_{Y|X}(y|x) \tag{30}$$

donc

$$f_{Y|X}(y|x) = f_Y(y) \tag{31}$$

et par conséquent

$$f_{XY}(x,y) = f_X(x)f_Y(y) \tag{32}$$

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

Indépendance et espérance

Theorem

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 thérome ainsi : Si X et Y sont indépendantes, alors

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

Indépendance et corrélation

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

⚠ Remarque : La réciproque est fausse : 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.

Loi des grands nombres I

Convergence en probabilité : Définition

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

$$\forall \epsilon > 0, \qquad \lim_{n \to \infty} \mathbb{P}\left(|X_n - X| \geq \epsilon\right) = 0. \tag{33}$$

Theorem (Loi des grands nombres)

Soit $X_1, \ldots, X_n$ 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 :

$$\forall \epsilon > 0, \qquad \lim_{n \to +\infty} \mathbb{P}\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mathbb{E}(X)\right| \geqslant \epsilon\right) = 0 \tag{34}$$

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

Loi des grands nombres II

⚠ Remarque : 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.

Vecteurs Aléatoires

- Variables aléatoires
- Couple de variables aléatoires
- Vecteurs Aléatoires
- Variables et vecteurs aléatoires gaussiens

Vecteurs Aléatoires

Definition

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

Fonction de répartition

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

$$F_{\boldsymbol{X}}(\mathbf{x}) = F_{\boldsymbol{X}}(x_1, \ldots, x_d) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_d \leq x_d) \qquad (35)$$

Vecteurs Aléatoires

f.d.p : cas continu

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

$$f_{\boldsymbol{X}}(\mathbf{x}) = f_{\boldsymbol{X}}(x_1, \ldots, x_p) = \frac{\partial^d F_{\boldsymbol{X}}(\mathbf{x})}{\partial x_1 \partial x_2 \ldots \partial x_d}. \tag{36}$$

Vecteurs Aléatoires

f.d.p : cas continu

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

$$f_{\boldsymbol{X}}(\mathbf{x}) = f_{\boldsymbol{X}}(x_1, \ldots, x_p) = \frac{\partial^d F_{\boldsymbol{X}}(\mathbf{x})}{\partial x_1 \partial x_2 \ldots \partial x_d}. \tag{36}$$

Si cette f.d.p existe, on alors

$$\mathbb{P}(\boldsymbol{X} \in A) = \int_A f_{\boldsymbol{X}}(\mathbf{x}) \, \mathrm{d}\mathbf{x}, \tag{37}$$

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

$$\int_{\mathbb{R}^d} f_{\boldsymbol{X}}(\mathbf{x}) \, \mathrm{d}\mathbf{x} = 1. \tag{38}$$

Loi de probabilité : cas discret

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

$$\mathbb{P}(\boldsymbol{X} \in A) = \sum_{\mathbf{x} \in A} P_{\boldsymbol{X}}(\mathbf{x}) \,, \tag{39}$$

et

$$\sum_{\mathbf{x} \in \mathcal{X}} P_{\boldsymbol{X}}(\mathbf{x}) = 1 \,, \tag{40}$$

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

Espérance et matrice de covariance d'un vecteur aléatoire

Soit $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ un vecteur aléatoire réel.

Espérance

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

$$\mathbb{E}[\boldsymbol{X}] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d])^T = (\mu_1, \ldots, \mu_d)^T \text{ avec } \mu_j = \mathbb{E}[X_j], j = 1, \ldots, d$$

Matrice de covariance

La matrice de variance-covariance (appelée aussi matrice de covariance) d'un vecteur aléatoire $\boldsymbol{X}$ est la matrice carrée parfois notée $\boldsymbol{\Sigma}$ dont le terme générique est donné par : $\boldsymbol{\Sigma}_{i,j} = \operatorname{cov}(X_i, X_j)$ Elle est définie comme :

$$\boldsymbol{\Sigma}_{\boldsymbol{X}} = \operatorname{cov}(\boldsymbol{X}) = \mathbb{E}\left[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^T\right] = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T] - \mathbb{E}[\boldsymbol{X}](\mathbb{E}[\boldsymbol{X}])^T, \tag{41}$$

$\mathbb{E}[\boldsymbol{X}]$ étant l'espérance mathématique de $\boldsymbol{X}$.

Espérance et matrice de covariance d'un vecteur aléatoire

En développant les termes on obtient la forme suivante :

$$\mathbf{\Sigma} = \begin{bmatrix} \mathrm{E}[(X_1-\mu_1)(X_1-\mu_1)] & \mathrm{E}[(X_1-\mu_1)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_1-\mu_1)(X_n-\mu_n)] \\ \mathrm{E}[(X_2-\mu_2)(X_1-\mu_1)] & \mathrm{E}[(X_2-\mu_2)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_2-\mu_2)(X_n-\mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n-\mu_n)(X_1-\mu_1)] & \mathrm{E}[(X_n-\mu_n)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_n-\mu_n)(X_n-\mu_n)] \end{bmatrix}$$

$$= \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1X_2) & \cdots & \mathrm{cov}(X_1X_p) \\ \mathrm{cov}(X_2X_1) & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_PX_1) & \cdots & \cdots & \mathrm{var}(X_p) \end{pmatrix} = \begin{pmatrix} \sigma^2_{X_1} & \sigma_{X_1X_2} & \cdots & \sigma_{X_1X_p} \\ \sigma_{X_2X_1} & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_pX_1} & \cdots & \cdots & \sigma^2_{X_p} \end{pmatrix}$$

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

Quelques propriétés de la matrice de covariance

La matrice de covariance possède les propriétés suivantes :

1. La matrice est **symétrique** car on a $cov(X, Y) = cov(Y, X)$
2. Elle est **semi-définie positive** (ses valeurs propres sont positives ou nulles).
3. Les éléments de sa diagonale ($\mathbf{\Sigma}_{i,i}$) représentent la variance de chaque variable, étant donné la propriété $cov(X, X) = var(X)$
4. Les éléments en dehors de la diagonale ($\mathbf{\Sigma}_{i,j}, i \neq j$) représentent la covariance entre les variables i et j.
5. la matrice de covariance d'un vecteur décorrélé ou indépendant est diagonale ($\mathbf{\Sigma}_{i,j} = 0 \ \forall i \neq j$)
6. etc

Matrice de corrélation

Definition

La matrice de corrélation (notée R) du vecteur aléatoire $\boldsymbol{X}$, définie de manière analogue à la matrice de covariance, est la matrice dont le terme général est le cœfficient de corrélation linéaire $\rho_{i,j}$ donné par l'équation (42).

Coœfficient de corrélation

Chaque terme $\rho_{i,j}$ représente la corrélation entre le couple de variables (X_i, X_j) du vecteur $\boldsymbol{X}$ et pour rappel est donné par

$$\rho_{i,j} = \frac{Cov(X_i, X_j)}{\sqrt{Var(X_i)Var(X_j)}} = \frac{\sigma_{i,j}}{\sigma_i \sigma_j}. \tag{42}$$

⚠ Remarque : 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.

Indépendance de vecteurs aléatoires

notion de VA i.i.d

Cas des variables discrètes

Soit $\boldsymbol{X} = (X_1, X_2, ..., X_n)$ une suite de variables aléatoires discrètes, et soit $(S_1, S_2, ..., S_n)$ une suite d'ensembles finis ou dénombrables tels que, pour tout $i \leq n$, $\mathbb{P}(X_i \in S_i) = 1$.

Definition

$(X_1, X_2, ..., X_n)$ est une suite de variables aléatoires indépendantes si et seulement si, pour tout $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \prod_{i=1}^{n} S_i$,

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

Indépendance de vecteurs aléatoires

Cas des variables aléatoires à densité

Soit une suite $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ de variables aléatoires réelles définies sur le même espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ et de densités de probabilité respectives f_i.

Definition

Les variables aléatoires réelles $(X_1, X_2, \ldots, X_n)$ sont dites indépendantes si et seulement si le vecteur $\boldsymbol{X}$ a une densité de probabilité f qui se définie par

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

Quelques statistique sur vecteurs aléatoires I

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

Vecteur moyen empirique

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_i \mathbf{x}_i \tag{43}$$

Matrice de covariance empirique

$$\mathbf{S} = \frac{1}{n} \sum_i (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \tag{44}$$

Variables et vecteurs aléatoires gaussiens

- Variables aléatoires
- Couple de variables aléatoires
- Vecteurs Aléatoires
- Variables et vecteurs aléatoires gaussiens

Introduction

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.

Variables et vecteurs aléatoires gaussiens

Definition

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 $f : \mathbb{R} \to \mathbb{R}^+$, telle que pour tout $x \in \mathbb{R}$:

$$f(x) = \frac{1}{\sigma\sqrt{2\,\pi}}\,\mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \tag{45}$$

Soit X une variable aléatoire réelle admettant pour densité de probabilité la loi normale (45), 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

Loi normale centrée réduite

La loi normale centrée réduite correspond à un cas particulier de la loi normale générale (45).

Definition

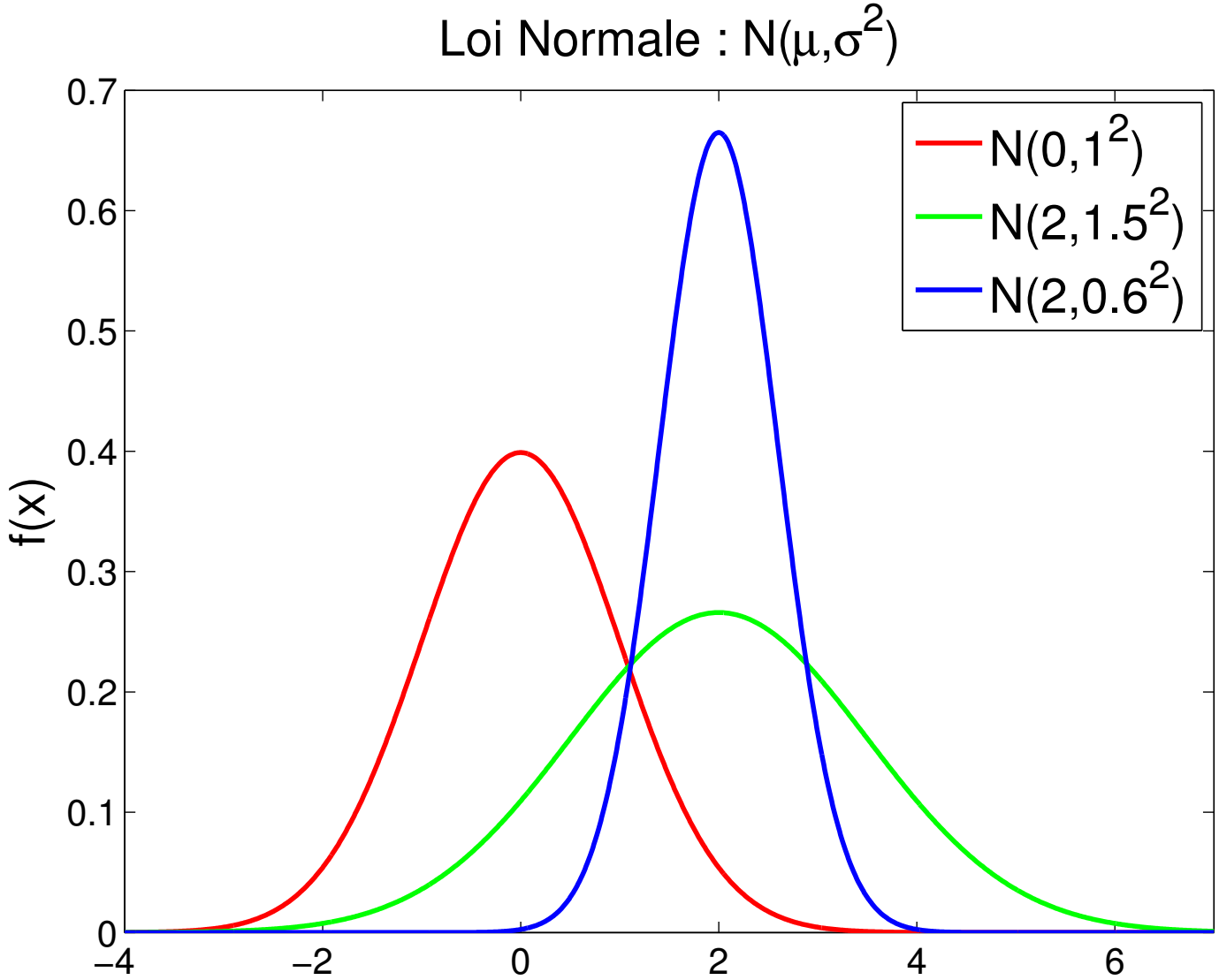
On appelle loi normale (ou gaussienne) centrée réduite la loi définie par la densité de probabilité $\varphi : \mathbb{R} \to \mathbb{R}^+$ définie par :

$$\varphi(x) = \frac{1}{\sqrt{2\,\pi}}\,\mathrm{e}^{-\frac{1}{2}x^2} \tag{46}$$

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

Représentations graphiques

Loi normale monodimensionnelle



Fonction de répartition de la loi normale

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.

Fonction de répartition 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 :

$$\Phi(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} \varphi(u)\,\mathrm{d}u = \int_{-\infty}^{x} \frac{1}{\sqrt{2\,\pi}}\,\mathrm{e}^{-\frac{u^2}{2}}\,\mathrm{d}u\,. \qquad (47)$$

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)

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-v^2} \tag{48}$$

par

$$\Phi(x) \;=\; \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right). \tag{49}$$

ou a encore

$$\operatorname{erf}(x) \;=\; 2\Phi\left(x\sqrt{2}\right) - 1. \tag{50}$$

Fonction de répartition de la loi normale générale

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

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right). \tag{51}$$

Quelques propriétés I

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

- son espérance et sa variance existent et $\mathbb{E}(X) = \mu,\ \mathrm{Var}(X) = \sigma^2 \geq 0$,
- la variable aléatoire $X^\star = \frac{X - \mathbb{E}(X)}{\sqrt{\mathrm{var}(X)}}$, c'est-à-dire $X^\star = \frac{X-\mu}{\sigma}$, suit la loi normale centrée réduite $\mathcal{N}(0, 1)$,
- si a, b sont deux réels ($a \neq 0$), alors la variable aléatoire $a\,X + b$ suit la loi normale $\mathcal{N}(a\,\mu + b,\ a^2\,\sigma^2)$

Théorème central limite

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

Theorem (Théorème Central Limite)

Soit $X_1, X_2, \cdots, X_n$ une suite de variables aléatoires réelles mutuellement indépendantes et suivant la meme densité f (variables i.i.d) d'espérance μ et d'écart-type σ. Soit la variable aléatoire somme

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

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

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

⚠ Remarque : : Ici on parle de convergence en probabilité.

Théorème central limite

$\Rightarrow$ 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.

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

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

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 converges vers la loi normale centrée réduite quand n tend vers l'infini.

Couple de variables aléatoires gaussiennes I

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 :

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

avec $\sigma_1 et \sigma_2 \geq 0$ 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.

Une corrélation nulle (décorrélation) implique l'indépendance quand les variables aléatoires sont gaussiennes.

Couple de variables aléatoires gaussiennes II

Démonstration.

Soit $\rho = 0$ dans l'équation (52). On a la densité jointe de $(X, Y$ est donnée par

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

qui est lé résultat recherché.

Vecteurs aléatoires gaussiens

Definition

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

Loi normale multidimensionnelle

La loi normale multidimensionnelle représente la distribution d'un vect. a. gaussien. Soit $\boldsymbol{X}$ un vect. a. gaussien de dimension d. On appelle loi normale multidimensionnelle d'espérance $\boldsymbol{\mu}$ et de matrice de covariance $\boldsymbol{\Sigma}$ la densité de probabilité $f : \mathbb{R}^d \to \mathbb{R}^+$ définie par :

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right), \qquad (52)$$

$|\boldsymbol{\Sigma}|$ étant le déterminant de $\boldsymbol{\Sigma}$.

Vecteurs aléatoires gaussiens

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

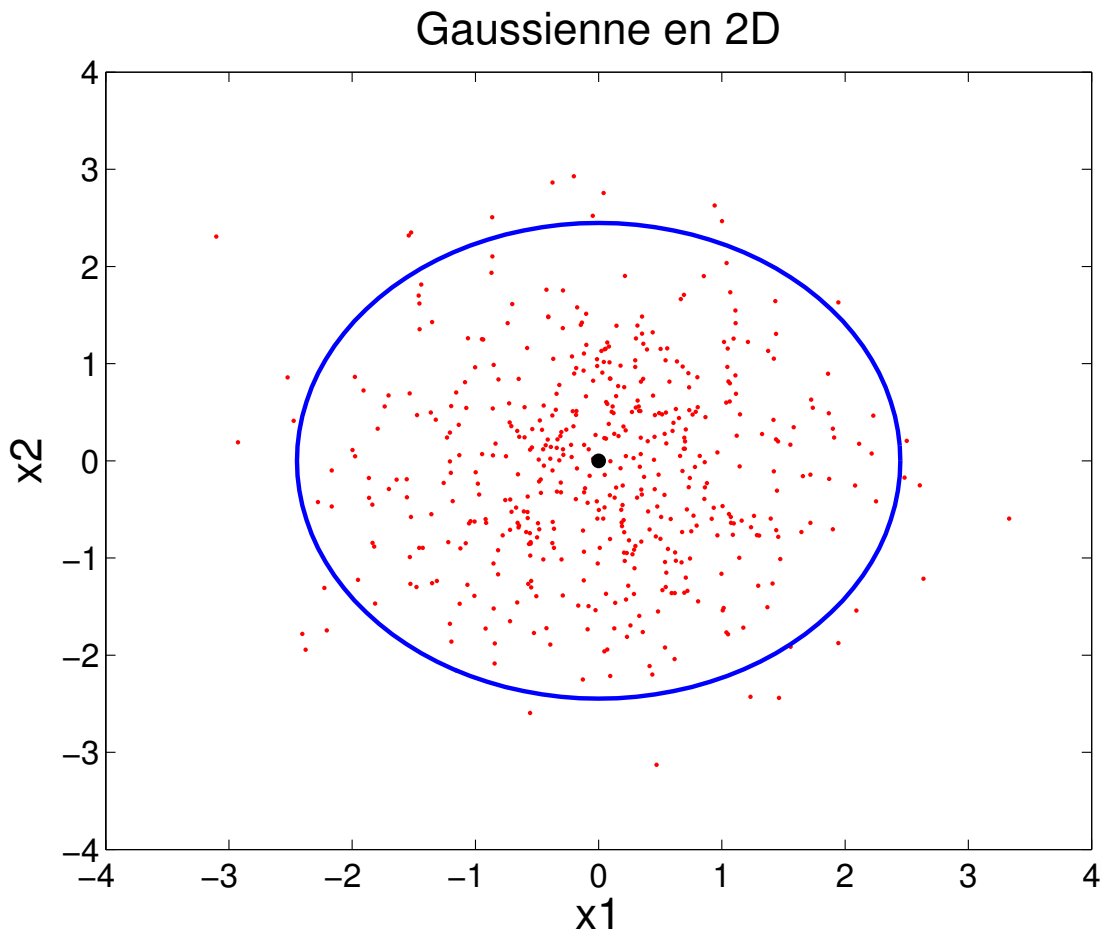
Distance de Mahalanobis

Le terme présent dans l'exponentielle dans (52) $(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})$ est le carré de la **distance de Mahalanobis**. Cette dernière est en effet donnée par

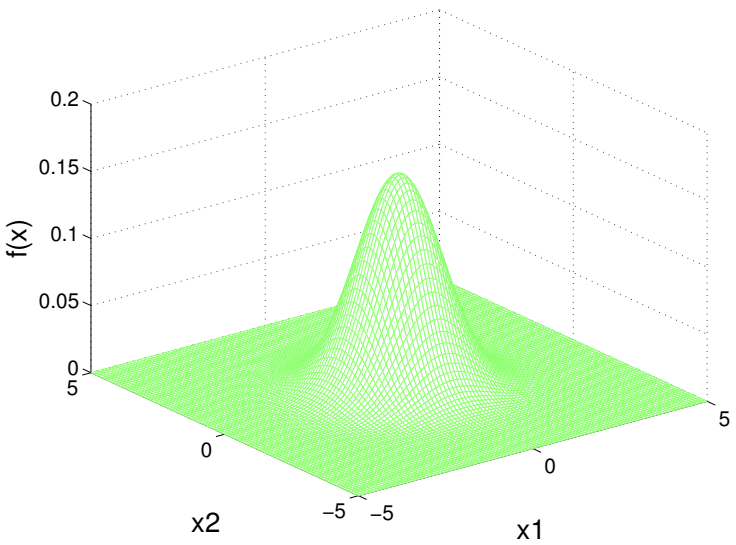
$$\sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}$$

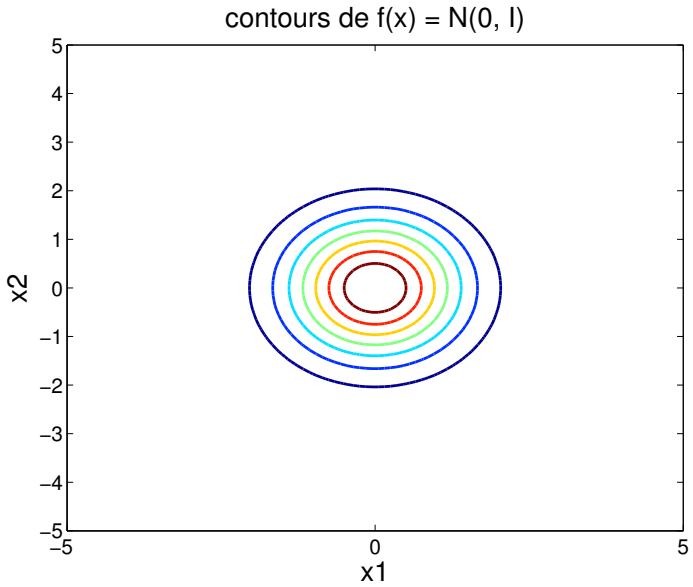
Représentations graphiques

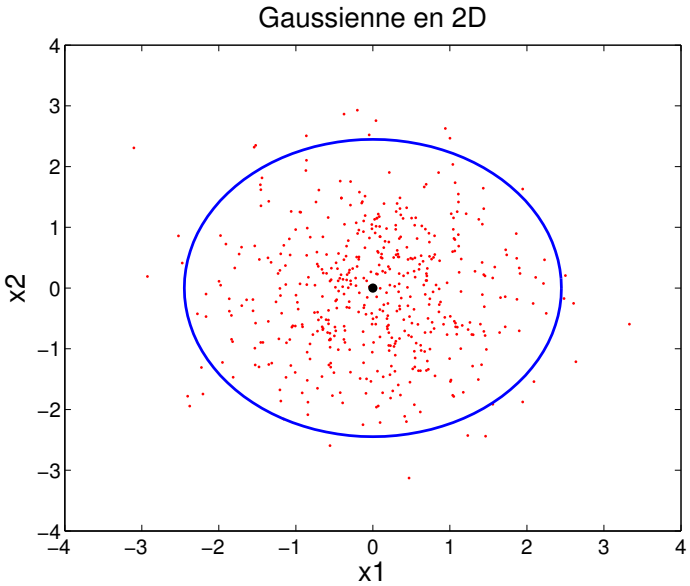
Loi normale multidimensionnelle

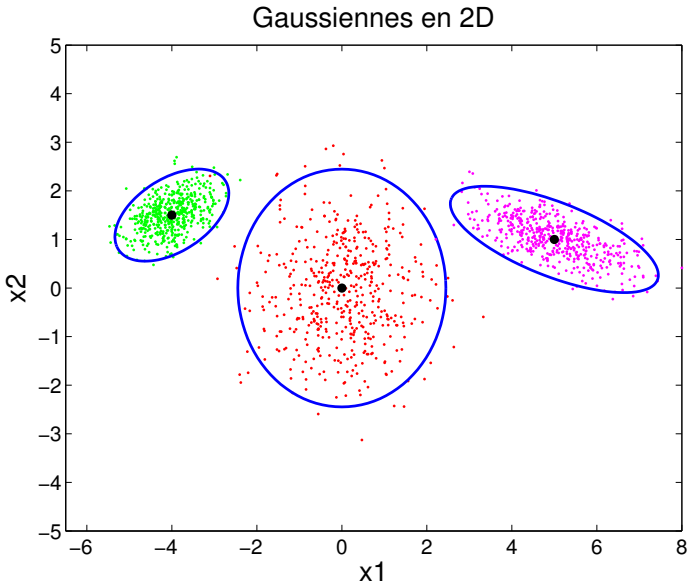


Repr. graphiques : Loi normale multidimensionnelle









Vecteurs aléatoires gaussiens

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

Queqlues Propriétés

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

Estimation de paramètres

- Introduction
- Critères de qualité pour les estimateurs
- Méthodes d'estimation

Estimation de paramètres

Pour étudier et caractériser un phénomène physique, naturel ou autre $\Rightarrow$ e.g., adoption d'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).

$\Rightarrow$ L'explication de ce phénomène nécessite l'estimation de ce modèle probabiliste à partir des données que l'on a observées (l'échantillon que l'on a à notre disposition).

$\Rightarrow$ Ceci consiste donc à estimer le(s) paramètre(s) θ de ce modèle à partir des données observées $(x_1, \ldots, x_n)$ (i.i.d 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.

Estimation de paramètres

Définition d'un estimateur

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

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

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

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

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

Critères de qualité pour les estimateurs

- Ce sont des 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.

$\Rightarrow$ Notre objectif est de déterminer un estimateur $\hat{\Theta} = h(X_1, \ldots, X_n)$ de θ.

$\Rightarrow$ Des propriétés comme la moyenne, la variance ou la distribution fournissent une mesure de qualité pour cet estimateur.

Critères de qualité pour les estimateurs

- Ce sont des 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.

$\Rightarrow$ Notre objectif est de déterminer un estimateur $\hat{\Theta} = h(X_1, \ldots, X_n)$ de θ.

$\Rightarrow$ Des propriétés comme la moyenne, la variance ou la distribution fournissent une mesure de qualité pour cet estimateur.

Estimateur vs Estimation

Une fois nous avons observé un échantillon de valeurs $(x_1, \ldots, x_n)$, la valeur de l'estimateur $\hat{\theta} = h(x_1, \ldots, x_n)$ qui est une valeur numérique, est appelé *estimation* du paramètre θ.

Absence de biais

Définition : **Absence de biais**

Un estimateur $\hat{\Theta}$ de θ est dis *sans biais* si

$$\mathbb{E}[\hat{\Theta}] = \theta, \tag{53}$$

$\Rightarrow$ en moyenne, on espère que $\hat{\Theta}$ est égal à la valeur du vrai paramètre θ.

Absence de biais

Définition : **Absence de biais**

Un estimateur $\hat{\Theta}$ de θ est dis *sans biais* si

$$\mathbb{E}[\hat{\Theta}] = \theta, \tag{53}$$

$\Rightarrow$ en moyenne, on espère que $\hat{\Theta}$ est égal à la valeur du vrai paramètre θ.

⚠ Remarque : Il est naturel que, si $\hat{\Theta}$ 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.

$\Rightarrow$ 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.

Variance minimale

Définition **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 $\Theta^\star$ de θ, à partir du même échantillon, on a :

$$\text{var}(\hat{\Theta}) \leq \text{var}(\Theta^\star). \tag{54}$$

Variance minimale

Définition **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 $\Theta^\star$ de θ, à partir du même échantillon, on a :

$$\text{var}(\hat{\Theta}) \leq \text{var}(\Theta^\star). \tag{54}$$

$\Rightarrow$ 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 estimations ont tendance à être plus proche de sa moyenne qui est la valeur du vrai paramètre.

Variance minimale

Définition **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 $\Theta^\star$ de θ, à partir du même échantillon, on a :

$$\text{var}(\hat{\Theta}) \leq \text{var}(\Theta^\star). \tag{54}$$

$\Rightarrow$ 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 estimations ont tendance à être plus proche de sa moyenne qui est la valeur du vrai paramètre.

$\Rightarrow$ 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 ? $\Rightarrow$ Thèorème : Borne de Cramer-Rao

Variance minimale : Borne de Cramer-Rao

Theorem (Borne de Cramer-Rao (Cramer-Rao Lower Bound (CRLB)))

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

$$\text{var}(\hat{\Theta}) \geq \left[n\mathbb{E}\left(\frac{\partial \ln f(X;\theta)}{\partial \theta} \right)^2 \right]^{-1} \tag{55}$$

⚠ Remarque : 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.

$\Rightarrow$ Cette inéquation fournit donc une borne inférieure de la variance de n'importe quel estimateur sans biais.

Variance minimale : Borne de Cramer-Rao

Information de Fisher

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 $\mathcal{I}_n(\theta)$.

Variance minimale : Borne de Cramer-Rao

Information de Fisher

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 $\mathcal{I}_n(\theta)$.

Borne de Cramer-Rao et Information de FIsher

La borne inférieure de Cramér-Rao alors se définie aussi par :

$$\text{var}(\hat{\Theta}) \geq \frac{1}{\mathcal{I}_n(\theta)} \tag{56}$$

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 : En anglais, la borne inférieure de Cramér-Rao s'appelle **Cramér-Rao Lower Bound** abréviée par CRLB.

Variance minimale : Borne de Cramér-Rao

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 (55) est équivalente à $-\left[\mathbb{E}\left(\frac{\partial^2 \ln f(X;\theta)}{\partial^2\theta}\right)\right]^{-1}$.

$\Rightarrow$ L'inégalité de Cramér-Rao peut également être alors mise sous la forme :

CRLB : deuxième forme opérationnelle

$$\text{var}(\hat{\Theta}) \geq -\left[n\mathbb{E}\left(\frac{\partial^2 \ln f(X;\theta)}{\partial^2\theta}\right)\right]^{-1}. \tag{57}$$

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

⚠ Remarque : : ces résultats concernent le cas d'un seul paramètre θ

$\Rightarrow$ Le résultat peut être facilement étendu au cas de plusieurs paramètres.

Variance minimale : Borne de Cramér-Rao

Cas de plusieurs paramètres : Soit $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)^T$ $(m \leq n)$ le vecteur des paramètres inconnus du modèle (la densité) $f(x; \theta_1, \cdots, \theta_m)$ pour lequel on cherche un estimateur $\hat{\boldsymbol{\Theta}} = (\hat{\Theta}_1, \cdots, \hat{\Theta}_m)^T$.

Variance minimale : Borne de Cramér-Rao

Cas de plusieurs paramètres : Soit $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)^T$ $(m \leq n)$ le vecteur des paramètres inconnus du modèle (la densité) $f(x; \theta_1, \cdots, \theta_m)$ pour lequel on cherche un estimateur $\hat{\boldsymbol{\Theta}} = (\hat{\Theta}_1, \cdots, \hat{\Theta}_m)^T$.

Borne de Cramér-Rao pour un vecteur paramètre

L'inégalité de Cramér-Rao, pour le cas de paramètres multiples, est de la forme

$$\text{cov}(\hat{\boldsymbol{\Theta}}) \geq \frac{\Lambda^{-1}}{n}, \tag{58}$$

ou le terme général de la matrice d'information de Fisher Λ est donné par :

$$\Lambda_{ij} = \Lambda(\theta_i, \theta_j) = \mathbb{E}\left[\left(\frac{\partial \ln f(X; \boldsymbol{\theta})}{\partial \theta_i}\right)\left(\frac{\partial \ln f(X; \boldsymbol{\theta})}{\partial \theta_j}\right)\right], \; i, j = 1, \cdots, m. \tag{59}$$

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

Variance minimale : Borne de Cramér-Rao

Transformée de la CRLB :

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

$$\text{CRLB pour var}(\hat{\mathbf{\Phi}}) = \left[\frac{\mathrm{d}\, g(\theta)}{\mathrm{d}\,\theta}\right]^2 \times \Big(\text{CRLB pour var}(\hat{\Theta})\Big) \qquad (60)$$

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

Efficacité d'un estimateur

Définition : **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 $\hat{\Theta}$

$$e(\hat{\Theta}) = \frac{\text{CRLB pour var}(\hat{\Theta})}{\text{var}(\hat{\Theta})} \tag{61}$$

$\Rightarrow$ L'efficacité d'un estimateur sans biais est ainsi inférieure ou égale à 1.

Estimateur efficace

Un estimateur sans biais ayant une efficacité égale à 1 est dit **efficace**.

Efficacité d'un estimateur

Définition : **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 $\hat{\Theta}$

$$e(\hat{\Theta}) = \frac{\text{CRLB pour var}(\hat{\Theta})}{\text{var}(\hat{\Theta})} \tag{61}$$

$\Rightarrow$ L'efficacité d'un estimateur sans biais est ainsi inférieure ou égale à 1.

Estimateur efficace

Un estimateur sans biais ayant une efficacité égale à 1 est dit **efficace**.

On souhaite aussi, en augmentant la taille de l'échantillon, pouvoir diminuer l'erreur d'estimation $\Rightarrow$ on parle de convergence

Consistance (ou convergence) d'un estimateur

Définition : **Consistance (ou convergence) d'un estimateur**

Un estimateur $\hat{\Theta}$ est dit **consistant** (on dit aussi convergent) pour θ si,

$$\lim_{n\to\infty} \mathbb{P}(|\hat{\Theta} - \theta| > \epsilon) = 0 \qquad \forall\, \epsilon > 0. \tag{62}$$

$\Rightarrow$ Convergence en probabilité

$\Rightarrow$ La probabilité de s'éloigner de la vraie valeur du paramètre de plus de ϵ tend vers 0 quand la taille de l'échantillon n augmente.

$\Rightarrow$ L'estimateur donc converge vers la valeur à estimer quand la taille de l'échantillon tends vers l'infini (asymptotiquement).

Consistance (ou convergence) d'un estimateur

Propriété

Un estimateur sans biais et de variance asymptotiquement nulle est convergent.
Soit $\hat{\Theta}$ un estimateur pour θ sur un échantillon de taille n. Alors, si

$$\lim_{n\to\infty} \mathbb{E}[\hat{\Theta}] = \theta, \quad \text{et} \quad \lim_{n\to\infty} \text{var}[\hat{\Theta}] = 0, \tag{63}$$

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

Suffisance d'un estimateur

Statistique suffisante (exhaustive)

Soit X un vecteur $i.i.d.$ de taille n. Soit θ un paramètre de la loi de probabilité des X_i. Une statistique $T(X)$ est dite exhaustive pour le paramètre θ (on dit aussi suffisante) si la probabilité conditionnelle d'observer X sachant $T(X)$ est indépendante de θ :

$$\mathbb{P}(X = x | T(X) = s, \theta) = \mathbb{P}(X = x | T(X) = s), \tag{64}$$

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

Statistique suffisante (exhaustive) et 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 u et v telles que :

$$f_\theta(x) = u(x)\, v(\theta, T(x))$$

Suffisance d'un estimateur

Définition : **Estimateur suffisant**

Soit $(X_1, X_2, \cdots, X_n)$ un échantillon $i.i.d.$ de X de distribution à paramètre θ. Si $Y = T(X_1, X_2, \cdots, X_n)$ est une statistique telle que, pour toute autre statistique $Z = u(X_1, X_2, \cdots, X_n)$, 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.

$\Rightarrow$ 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.

Suffisance et critère de factorisation de Fisher-Neyman

Si une statistique suffisante pour un paramètre θ existe, le théorème 82 suivant, fournit un moyen de la trouver.

Critère de factorisation de Fisher-Neyman

Soit $Y = T(X_1, \ldots, X_n)$ 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 :

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta) = u\left(T(x_1, \ldots, x_n), \theta\right) v(x_1, \ldots, x_n). \quad (65)$$

Dans le cas discret on a :

$$P(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} P(x_i; \theta) = v\left(T(x_1, \ldots, x_n), \theta\right) v(x_1, \ldots, x_n). \quad (66)$$

Suffisance et critère de factorisation de Fisher-Neyman

Le résultat précédent peut être étendu au cas de paramètres multiples.

Soit $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)^T$, $m \leq n$ le vecteur paramètre. Alors $Y_1 = T(X_1, \ldots, X_n), \ldots, Y_r = T(X_1, \ldots, X_n)$, $r \geq m$ est un ensemble de statistique suffisantes pour $\boldsymbol{\theta}$ si et seulement si

$$f(x_1, \ldots, x_n; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) = u\left(\boldsymbol{T}(x_1, \ldots, x_n), \boldsymbol{\theta}\right) v(x_1, \ldots, x_n). \quad (67)$$

avec $\boldsymbol{T} = (T_1, \ldots, T_r)^T$.

L'expression dans le cas discret est similaire.

Méthodes d'estimation I

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**.

1. 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.
2. 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.

Dans ce cours, nous verrons en particulier les méthodes du maximum de vraisemblance (et moindres carées) (estimation ponctuelle) pour leur usage très commun en estimation de modèles probabilistes et l'estimation par intervalle de confiance

Méthode du maximum de vraisemblance (Maximum Likelihood)

- Définition de la fonction de vraisemblance
- Maximum de vraisemblance
- Propriétés
- Cas gaussien

Méthode du maximum de vraisemblance I

Introduction

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.

Fonction de vraisemblance

Définition : **Fonction de vraisemblance**

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

$$L(\theta;\boldsymbol{x}) = L(\theta; x_1, \ldots, x_n) = f(x_1, \ldots, x_n; \theta). \tag{68}$$

Fonction de vraisemblance

Définition : **Fonction de vraisemblance**

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

$$L(\theta;\boldsymbol{x}) = L(\theta; x_1, \ldots, x_n) = f(x_1, \ldots, x_n; \theta). \tag{68}$$

Vraisemblance pour un échantillon *i.i.d.*

Pour le cas *i.i.d.*, la fonction de vraisemblance est donnée par

$$\begin{aligned} L(\theta;\boldsymbol{x}) = L(\theta; x_1, \ldots, x_n) &= f(x_1;\theta) f(x_2;\theta) \cdots f(x_n;\theta) \\ &= \prod_{i=1}^{n} f(x_i;\theta). \end{aligned} \tag{69}$$

Fonction de vraisemblance

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

$$\begin{aligned} L(\theta; \boldsymbol{x}) &= P(x_1;\theta)P(x_2;\theta)\cdots P(x_n;\theta) \\ &= \prod_{i=1}^{n} P(x_i;\theta). \end{aligned} \tag{70}$$

⚠ Remarque : On peut aussi rencontrer la notation $L(\theta)$ de la vraisemblance de θ au lieu de $L(\theta; x_1, \ldots, x_n)$ (notamment dans ce cours :)).

$\rightarrow$ pour des valeurs d'échantillon données, la fonction de vraisemblance est seulement fonction du paramètre θ.

Maximum de vraisemblance

Définition : **Définition du maximum de vraisemblance**

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 $(x_1, \ldots, x_n)$ viennent de la population de densité $f(x; \theta)$.

Maximum de vraisemblance : Cas d'un seul paramètre θ

maximum(s) d'une fonction

Mathématiquement, le maximum d'une fonction $L(\theta)$ correspond à la valeur de θ pour laquelle la dérivée de L par rapport à θ est nulle :

$$\frac{\mathrm{d}\, L(\theta)}{\mathrm{d}\, \theta} = 0$$

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 maximum vérifie la dérivée seconde par rapport à θ est négative :

$$\frac{\mathrm{d}^2\, L(\theta)}{\mathrm{d}^2\, \theta} < 0$$

Maximum de vraisemblance : Cas d'un seul paramètre θ

Estimateur du Maximum de vraisemblance (MV)

L'*estimateur du maximum de vraisemblance* (MV) de θ, noté $\hat{\theta}$, à partir des valeurs de l'échantillon $(x_1, \cdots, x_n)$ peut être déterminé à partir de

$$\frac{\mathrm{d}\, L(\theta; x_1, \ldots, x_n)}{\mathrm{d}\,\theta}|_{\theta=\hat{\theta}} = 0. \tag{71}$$

$$\frac{\mathrm{d}^2\, L(\theta; x_1, \ldots, x_n)}{\mathrm{d}^2\,\theta}|_{\theta=\hat{\theta}} < 0 \tag{72}$$

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

Maximum de vraisemblance

⚠ Remarque :

Bien que la plupart des vraisemblances soient différentiables, les solutions de l'équation de vraisemblance (71) ne s'expriment pas toujours par des formes analytiques.

$\Rightarrow$ 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)

$\Rightarrow$ par exemple la montée de gradient, l'algorithme de Newton Raphson, l'algorithme EM, etc.

Maximum de vraisemblance

le logarithme est une fonction monotone et la fonction de vraisemblance étant positive

$\Rightarrow$ la fonction de vraisemblance atteint donc son maximum pour la même valeur que son logarithme

Fonction de **log-vraisemblance**

Maximiser la fonction de vraisemblance revient à maximiser 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.

Maximum de vraisemblance

L'équation de vraisemblance devient donc :

$$\frac{\mathrm{d} \ln L(\theta; x_1, \ldots, x_n)}{\mathrm{d}\,\theta}|_{\theta=\hat{\theta}} = 0. \tag{73}$$

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**

⚠ Remarque : (lorsque tous les maxima ont été identifiés, seul le plus grand d'entre-eux doit être retenu).

Maximum de vraisemblance : Vraisemblance multivariée I

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.

Fonction de log-vraisemblance :

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

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

$$\frac{\mathrm{d} \ln L(\theta_1, \ldots, \theta_m; x_1, \ldots, x_n)}{\mathrm{d}\, \theta_j}|_{\theta_j = \hat{\theta}_j} = 0 \quad \text{pour } j = 1, \ldots, m \qquad (74)$$

Maximum de vraisemblance : Vraisemblance multivariée II

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

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

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

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

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

Propriétés du maximum de vraisemblance

Soit $\hat{\theta}$ la valeur de l'estimateur du maximum de vraisemblance $\hat{\Theta}$ de θ estimée à partir de l'échantillon $(x_1, \cdots, x_n)$ de taille n

Convergence

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

Absence de biais et efficacité asymptotiques

Quand n tend vers l'infini on a :

$$\lim_{n\to\infty} \mathbb{E}[\hat{\Theta}] = \theta \quad : \text{asymptotiquement sans biais} \tag{75}$$

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

Des résultats analogues sont obtenus lorsque X est une v.a. discrète.

Propriétés du maximum de vraisemblance

Normalité asymptotique

La distribution de $\hat{\Theta}$ tend vers une distribution normale lorsque n devient grand. L'EMV est donc *asymptotiquement normal*.

$$\sqrt{n}\,(\hat{\Theta} - \theta) \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}\left(0, \mathcal{I}_n(\theta)^{-1}\right). \tag{76}$$

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 $g(\hat{\Theta})$.

$\Rightarrow$ 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 $\hat{\sigma}^2$.

Maximum de vraisemblance dans le cas Gaussien

Soit $(X_1, \ldots, X_n)$ un échantillon de variables aléatoires réelles issues d'une population de densité normale $\mathcal{N}(\mu, \sigma^2)$, alors

1. la moyenne empirique $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur sans biais de l'espérance μ
2. la variance empirique *corrigée* $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_i)^2$ est un estimateur sans biais de la variance σ^2

et on a

$$\begin{aligned} \bar{X} &\sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \\ \frac{(n-1)S^2}{\sigma^2} &\sim \chi^2_{n-1} \end{aligned}$$

Loi de student

Soit Z une v.a de loi normale centrée et réduite et soit Uune v.a indépendante de Z et distribuée suivant la loi du χ^2 à n degrés de liberté. Par définition la variable

$$T = \frac{Z}{\sqrt{U/n}}$$

suit une loi de Student à n degrés de liberté.

Loi du χ^2

Soient $X_1, \ldots, X_n$, n v.a. indépendantes suivant des lois normales de moyennes respectives μ_i et d'écart-type σ_i ; $Y_i = \frac{X_i - \mu_i}{\sigma_i}$ leurs variables centrées et réduites, alors par définition la variable X , telle que

$$X := \sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{k} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2$$

suit une loi du χ^2 à n degrés de liberté.

Maximum de vraisemblance dans le cas Gaussien

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}.

Méthode des Moindres Carrées (Least Squares)

- Définition des Moindres Carrés
- Moindres Carrés
- Propriétés de l'estimateur des moindres carrés

Méthode des Moindres Carrées I

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) :

$$Y = f(X) + E \tag{77}$$

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

Méthode des Moindres Carrées II

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

Critère des moindres carrés

Soit le modèle

$$Y_i = f(X_i) + E_i \tag{78}$$

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

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

Définition : **Critères des moindres carrés**

L'erreur quadratique est donnée par la somme des carrés des résidus (Residual Sum of Squares (RSS)) :

$$\text{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2 . \tag{79}$$

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 qudratique dans ce cas est donnée par

$$RSS(\theta) = \sum_{i=1}^{n} (y_i - f(x_i;\theta))^2 . \tag{80}$$

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 qudratique dans ce cas est donnée par

$$RSS(\theta) = \sum_{i=1}^{n} (y_i - f(x_i;\theta))^2 . \tag{80}$$

Définition : **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(θ).

$$\hat{\theta} = \arg\min_{\theta} \text{RSS}(\theta) \tag{81}$$

En effet, en choisissant une valeur de θ qui minimise RSS(θ), 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 $((x_1, y_1), \ldots, (x_n, y_n))$.

Moindres carrés : Cas d'un seul paramètre θ

Estimateur des moindres carrés (MC)

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

$$\frac{\mathrm{d}\,\mathrm{RSS}(\theta)}{\mathrm{d}\,\theta}|_{\theta=\hat{\theta}} = 0. \tag{82}$$

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

$$\frac{\mathrm{d}^2\,\mathrm{RSS}(\theta)}{\mathrm{d}^2\,\theta}|_{\theta=\hat{\theta}} > 0 \tag{83}$$

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

Moindres Carrés

⚠ Remarque :

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

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

$\Rightarrow$ 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

Moindres Carrés : cas de paramètres multiples I

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

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

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

$$\frac{\partial \text{RSS}(\boldsymbol{\theta})}{\partial \theta_j}|_{\theta_j = \hat{\theta}_j} = 0 \quad \text{pour} \, j = 1, \ldots, m \tag{84}$$

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 $((x_1, y_1), \cdots, (x_1, y_n))$ de taille n

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 bias

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 $((x_1, y_1), \cdots, (x_1, y_n))$ de taille n

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 bias

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.

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 $((x_1, y_1), \cdots, (x_1, y_n))$ de taille n

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 bias

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 ($e_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$) :

Normalité

La distribution de $\hat{\Theta}$ est normale centrée sur le vrai paramètre θ

Régression linéaire

- Introduction
- Le modèle linéaire simple
- Estimation par moindres carrés
- Formulation vectorielle

Régression linéaire I

L'estimation de paramètres et l'évaluation de la qualité d'un estimateur seront appliqués dans se chapitre à une famille de problèmes très connues en statistique et estimation qui est celle de la régression simple et multiple.

Une situation qu'on rencontre couramment est celle dans laquelle une variable aléatoire Y est fonction d'une ou plusieurs variables indépendantes (déterministes) $(x_1, \ldots, x_m)$.

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.

Régression linéaire II

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 $(x_1, \ldots, x_m)$ étant donnés un échantillon des couples $((\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n))$ des variables

$(Y_1, \ldots, Y_n)$ de la variable Y et les valeurs associées des variables explicatives $(\mathbf{x}_1, \ldots, \mathbf{x}_m)$, telle que x_j, $j = 1, \ldots, m$ pour chaque valeurs observée de Y_i, $i = 1 \ldots, n$.

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

Le modèle linéaire simple I

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

$$Y = \beta_0 + \beta_1 x + \epsilon, \tag{85}$$

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.

Le modèle linéaire simple I

Le bruit est généralement supposé d'espérance nulle et de 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 dont 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 $((x_1, Y_1), \ldots, (x_n, Y_n))$ [2].

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.

Le modèle linéaire simple I

Le modèle s'écrit donc sous la forme

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{86}$$

où ϵ_i est le bruit associée à la $i$ème variable aléatoire. Étant donnés 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 :

$$y_i = \beta_0 + \beta_1 x_i + e_i, \tag{87}$$

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 (par exemple par maximum de vraisemblance ou encore par inférence bayésienne (maximum a posteriori)).

Le modèle linéaire simple : Estimation par moindres carrés

La méthode des moindres carrés est une approche d'estimation ponctuelle des paramètres de régression (β_0, β_1).

fournit les estimations $(\hat{\beta}_0, \hat{\beta}_1)$ qui minimisent la somme des écarts quadratiques entre les valeurs observées y_i et l'espérance $\beta_0 + \beta_1 x_i$ du modèle de Y_i.

D'après (87), l'écart entre la valeur d'une observation et l'espérance du modèle est

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

La somme des carrés des résidus est donc donnée par

$$\text{RSS}(\beta_0, \beta_1) = Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2. \qquad (88)$$

RSS(β_0, β_1) est quadratique $\Rightarrow$ minimisation analytique

Le modèle linéaire simple : Estimation par moindres carrés

Theorem

Considérons le modèle de régression simple donné par l'équation (85). Soit $((x_1, y_1), \ldots, (x_n, y_n))$ *un échantillon de valeurs observées de* Y *et leurs veleurs associées de* x*. Alors les estimations des moindres carrées ordinaires (MCO) de* β_0 *et* β_1 *sont données par :*

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{89}$$

$$\hat{\beta}_1 = \left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right]\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{-1}, \tag{90}$$

où $\bar{x}$ *représente la moyenne empirique des* x_i *et* $\bar{y}$ *la moyenne moyenne empirique es* y_i *et sont donnés par* $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ *et* $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ *.*

Le modèle linéaire simple : Estimation par moindres carrés

Preuve. Selon les MCO, les estimations $\hat{\beta}_0$ et $\hat{\beta}_1$ sont celles qui minimisent la somme des carrées des résidus (88) par rapport à $\hat{\beta}_0$ et $\hat{\beta}_1$.
Mathématiquement on note

$$\begin{aligned}(\hat{\beta}_0, \hat{\beta}_1) &= \arg \min_{(\beta_0,\beta_1)\in\mathbb{R}^2} \text{RSS}(\beta_0, \beta_1) \\ &= \arg \min_{(\beta_0,\beta_1)\in\mathbb{R}^2} \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2. \qquad (91)\end{aligned}$$

Nous aurons donc déterminé les estimations $(\hat{\beta}_0, \hat{\beta}_1)$. Ainsi, nous avons :

$$\frac{\partial Q}{\partial \beta_0} = \frac{\partial \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2}{\partial \beta_0} = -2 \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)$$

$$\frac{\partial Q}{\partial \beta_1} = \frac{\partial \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i \big(y_i - (\beta_0 + \beta_1 x_i)\big) \cdot$$

Le modèle linéaire simple : Estimation par moindres carrés

Ensuite, en annulant ces dérivés nous avons :

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

ce qui donne les **équations normales**

$$n\hat{\beta}_0 = \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i \quad (92)$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \quad (93)$$

Le modèle linéaire simple : Estimation par moindres carrés

1. En divisant par n la première équation on obtient la valeur de $\hat{\beta}_0$

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

2. 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

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

Le modèle linéaire simple : Estimation par moindres carrés

Il reste à vérifier si la dérivée partielle seconde de Q par rapport à au moins l'un des paramètres est positive et que le déterminant de la matrice des dérivées partielles secondes de Q est strictement positif.

- On a

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

- Le déterminant de la matrice des dérivées partielles secondes de Q est :

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

Notez que ce déterminant est nul si tous les x_i prennent la même valeur.
$\Rightarrow$Au moins deux valeurs x_i distinctes sont nécessaires pour la détermination de (β_0, β_1).

Régression linéaire : 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 $(y_1, \ldots, y_n)$ l'ensemble des valeurs observées de la variable dépendante Y et $(x_1, \ldots, x_n)$ l'ensemble des valeurs observées de la variable explicative x.

⚠ Remarque : L'ensemble des couples $\big((x_1, y_1), \ldots, (x_n, y_n)\big)$ 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

Régression linéaire : Formulation vectorielle

Selon le modèle de régression linéaire simple on a

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + e_1, \\ y_2 &= \beta_0 + \beta_1 x_2 + e_2, \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_i + e_n \cdot \end{aligned} \tag{94}$$

Régression linéaire : Formulation vectorielle

Selon le modèle de régression linéaire simple on a

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + e_1, \\ y_2 &= \beta_0 + \beta_1 x_2 + e_2, \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_i + e_n \cdot \end{aligned} \tag{94}$$

Notons par

- $\mathbf{y} = (y_1, \ldots, y_n)^T$ le vecteur des valeurs d'observations de Y,
- $\mathbf{e} = (e_1, \ldots, e_n)^T$ le vecteur des valeurs des résidus
- $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ le vecteur des coefficients de régressions à estimer
- $\mathbf{X}$ la *matrice de régression* (matrice de *design* ou de *Vendermonde*)

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

Régression linéaire : Formulation vectorielle

le modèle (94) s'écrit donc sous la forme matricielle suivante :

Régression linéaire : Formulation vectorielle

$$\boldsymbol{y} \quad = \quad \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \cdot \tag{95}$$

La somme des écarts quadratiques $\sum_{i=1}^{n} e_i^2$ est maintenant donnée par :

Régression linéaire : Formulation vectorielle

$$\mathrm{RSS}(\boldsymbol{\beta}) = \mathbf{e}^T\mathbf{e} \quad = \quad (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{96}$$

Régression linéaire : Formulation vectorielle

L'estimation $\hat{\boldsymbol{\beta}}$ par moindres carrés de $\boldsymbol{\beta}$ s'obtient en minimisant (96) qui est une fonction quadratique en $\boldsymbol{\beta}$.

On a :

- $\text{RSS}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$
- $\frac{\partial \text{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$
- $\frac{\partial \text{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = \mathbf{0} \Rightarrow -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}$

$\Rightarrow$ les équations normales

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}\,. \tag{97}$$

Régression linéaire : Formulation vectorielle

$$\begin{aligned} \mathbf{X}^T\mathbf{X}\hat{\beta} &= \mathbf{X}^T\mathbf{y} \\ (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ \hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \end{aligned}$$

$\Rightarrow$ Estimateur des moindres carrés :

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\,. \tag{98}$$

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

Estimation par intervalle

- Estimation par intervalle
- Intervalle de confiance
- Calcul d'un intervalle de confiance
- Loi normale : Intervalle de confiance sur μ
- Loi normale : Intervalle de confiance sur σ^2

Estimation par intervalle I

Nous allons maintenant voir une autre approche d'estimation des paramètres **l'estimation par intervalle**.

Lorsqu'on est intéressé non seulement par l'estimation en elle-même mais aussi par le niveau de confiance et la marge d'erreur, on effectue une 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.

Estimation par intervalle II

L'intervalle est appelé **intervalle de confiance** (IC) le niveau de confiance est aussi appelé **précision** ou **cœfficient de confiance**.

Intervalle de confiance I

Définition : **Intervalle de confiance**

Soit $(X_1, \ldots, X_n)$ un échantillon de variables aléatoires issues d'une population de densité $f(x;\theta)$, θ étant le (vecteur) paramètre à estimer. Supposons aussi que $T_1(X_1, \ldots, X_n)$ et $T_2(X_1, \ldots, X_n)$ 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

$$\mathbb{P}(T_1 < \theta < T_2) = 1 - \alpha \cdot \tag{99}$$

α représente le *risque* que le vrai paramètre θ ne soit pas dans cet intervalle et $1-\alpha$ s'appelle niveau ou (cœfficient) de confiance. Une estimation par intervalle de confiance sera donc d'autant meilleure que l'intervalle sera petit pour un cœfficient 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.

Intervalle de confiance II

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.

Intervalle de confiance III

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

$$\mathbb{P}(a < \theta < b) = 1 - \alpha \Longrightarrow \mathbb{P}(\theta < a) + \mathbb{P}(\theta > b) = \alpha \qquad (100)$$

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.

Loi normale : Intervalle de confiance pour μ avec σ connu I

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

$$U = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \quad \sim \quad \mathcal{N}(0,1) \tag{101}$$

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

$$\mathbb{P}(U \geq \frac{\alpha}{2}) = \alpha - \mathbb{P}(U \leq \frac{\alpha}{2}) = \frac{\alpha}{2} \tag{102}$$

La notion de quantile est définie de la façon suivante :

Loi normale : Intervalle de confiance pour μ avec σ connu II

Definition

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

$$\mathbb{P}(U \leq q_\alpha) = \alpha \tag{103}$$

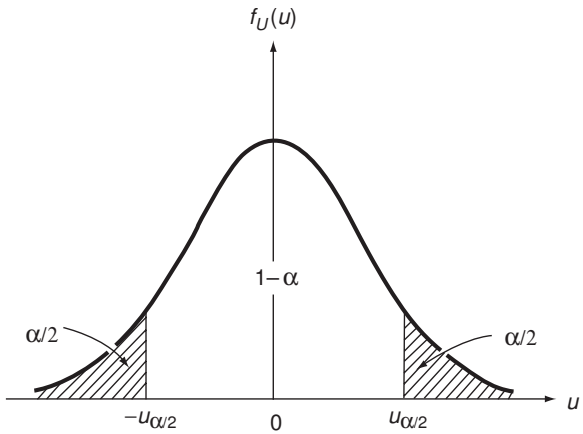


Figure – Loi normale centrée réduite et IC à $[100(1-\alpha)]\%$

Loi normale : Intervalle de confiance pour μ avec σ connu III

⚠ 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^n_α pour la loi χ^2_n, etc

Le risque étant symétrique, d'après (102) on a

$$\mathbb{P}(-u_{\frac{\alpha}{2}} \leq U \leq u_{\frac{\alpha}{2}}) = 1 - \alpha \tag{104}$$

et d'après (101) on obtient

$$\mathbb{P}\left(\bar{X} - u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \tag{105}$$

d'où l'intervalle de confiance sur μ :

$$\mathrm{IC}_{1-\alpha}(\mu) = \left[\bar{X} - u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\,,\, \bar{X} + u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] \tag{106}$$

Loi normale : Intervalle de confiance pour μ avec σ connu IV

exemple pour $\alpha = 0.05$

$$\mathrm{IC}_{0.95}(\mu) = \left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}\,,\, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right] \tag{107}$$

Loi normale : Intervalle de confiance pour μ avec σ inconnu I

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 a 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

$$T = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \tag{108}$$

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

la statistique (108) pour calculer l'IC s'écrit dont

$$T = \frac{U}{\sqrt{\frac{Z}{n-1}}} \sim \text{ Loi de Student à n-1 degrés de liberté} \tag{109}$$

Loi normale : Intervalle de confiance pour μ avec σ inconnu II

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.

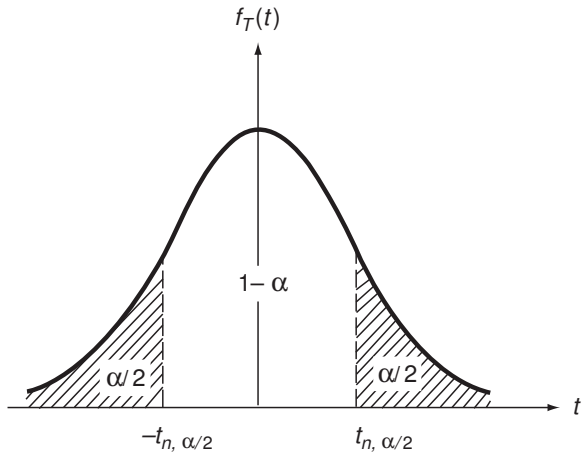


Figure – Loi de Student à $n-1$ degrés de liberté et IC à $[100(1-\alpha)]\%$

Loi normale : Intervalle de confiance pour μ avec σ inconnu III

L'intervalle de confiance est donc donné par

$$\mathbb{P}(-t_{n-1,\frac{\alpha}{2}} \leq T \leq t_{n-1,\frac{\alpha}{2}}) = 1-\alpha \tag{110}$$

On obtient donc l'intervalle de confiance comme précédemment

$$\mathrm{IC}_{1-\alpha}(\mu) = \left[\bar{X} - t_{n-1,\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\,,\, \bar{X} + t_{n-1,\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right] \tag{111}$$

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 : 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.

Loi normale : Intervalle de confiance pour σ lorsque μ connu I

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 :

$$D = \frac{n}{\sigma^2} V^2 \sim \chi^2_n \tag{112}$$

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

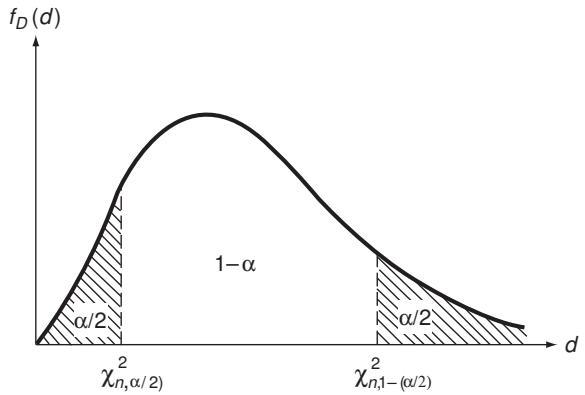


Figure – Loi de χ^2 à n degrés de liberté et IC à $[100(1-\alpha)]\%$

Loi normale : Intervalle de confiance pour σ lorsque μ connu II

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

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

donc

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

finalement on obtient

$$IC_{1-\alpha}(\sigma^2) = \left[\frac{nV^2}{\chi^2_{n,1-\frac{\alpha}{2}}}, \leq \frac{nV^2}{\chi^2_{n,\frac{\alpha}{2}}}\right] \tag{113}$$

Loi normale : Intervalle de confiance pour σ lorsque μ inconnu I

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 :

$$D = \frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1} \tag{114}$$

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

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

donc

Loi normale : Intervalle de confiance pour σ lorsque μ inconnu II

$$\mathrm{IC}_{1-\alpha}(\sigma^2) = \left[\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} , \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}} \right] \tag{115}$$

⚠ 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

Tests d'hypothèses

- Région de rejet d'un test
- Erreurs associées à un test
- Statistiques de test

Tests d'hypothèses

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.

$\Rightarrow$ 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

$$H_0 : \theta \in \Theta_0 \tag{116}$$

où Θ_0 est l'ensemble de valeurs supposée du paramètre θ selon H_0.

Tests d'hypothèses

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

$$H_1 : \theta \in \Theta_1 \tag{117}$$

avec Θ_1 un sous ensemble de l'ensemble des paramètres disjoint de θ_0.

Tests d'hypothèses

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**.

Région de rejet d'un test et erreurs

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 sou la forme suivante

$$R = \{x : T(x) > s\} \tag{118}$$

où T est une **statistique de test** et s un **seuil**.

$\Rightarrow$ 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

Région de rejet d'un test et erreurs

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).

Erreur de première espèce

L'erreur de première espèce correspond au cas où l'on rejette H_0 (décider H_1) alors que celle-ci est vraie.

Erreur de deuxième espèce

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.

Erreurs et risques suite à un test

Les décisions possibles sont récapitulées par le tableau suivant.

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

Table – Récapitulatif des décisions en test d'hypothèse

Pour chacune des deux erreurs, on associe un une probabilité (**un risque**).

Définition : risque de première espèce

Le risque de première espèce est notée 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%.

Définition : risque de deuxième espèce

Le risque de deuxième espèce représente quant à lui le risque de retenir H_0 à tort. Il est noté β.

Erreurs et risques suite à un test

Définition : niveau de confiance du test

Le niveau de confiance du test est donc $1-\alpha$ qui correspond à retenir H_0 à raison.

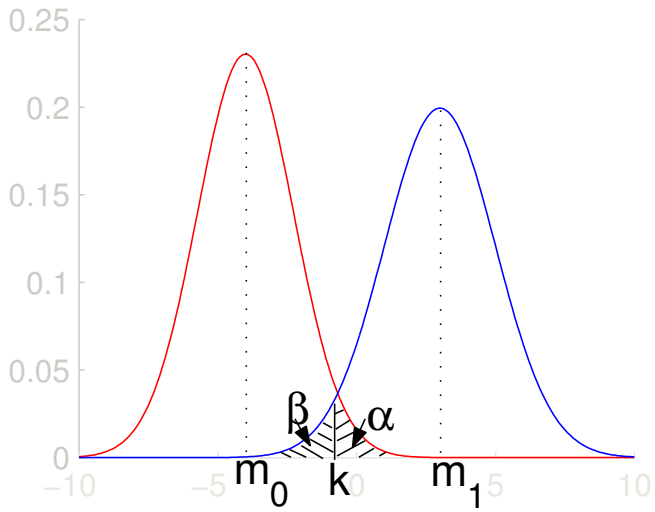
Définition : puissance d'un test

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


0.25
0.2
0.15
0.1
0.05
0
−10
−5
0
5
10
β
α
m0
k
m1

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.

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

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

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

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

Test du rapport de vraisemblance (ou Test de Neyman-Pearson) I

Exemple :

Soit un échantillon d'observations $i.i.d.$ $(x_1, \ldots, x_n)$ 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 $\mu < \mu_1$

Test du rapport de vraisemblance (ou Test de Neyman Pearson) II

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

$$\begin{aligned} L(\mu; \mathbf{x}) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\,\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right] \\ &= \frac{1}{(\sigma^2\sqrt{2\,\pi})^n} \exp\left[-\sum_{i=1}^{n} \frac{1}{2\sigma}\left(x_i - \mu\right)^2\right] \end{aligned} \tag{119}$$

le rapport de vraisemblance est don donné par

$$\frac{L(\mu_1; \mathbf{x})}{L(\mu_0; \mathbf{x})} = \exp\left[\frac{1}{2\sigma^2} 2\left(\mu_1 - \mu_0\right) \sum_{i=1}^{n} x_i - \frac{n}{2\sigma^2}\left(\mu_1^2 - \mu_0^2\right)\right] \tag{120}$$

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

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

Test du rapport de vraisemblance (ou Test de Neyman-Pearson) III

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

$$R = \{\mathbf{x} : \bar{x} > \mu_0 + u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\} \tag{121}$$

Test de Wald I

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 : $\frac{\sqrt{n}(\widehat{\Theta}-\theta_0)}{\sqrt{\text{var}(\hat{\Theta})}} \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,1)$.

Dans le *test de Wald*, la statistique de test est donnée par

$$\frac{\widehat{\Theta} - \theta_0}{\sqrt{\text{var}(\hat{\Theta})}} \tag{122}$$

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

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

Test de Wald II

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

$$U = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \tag{123}$$

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

Test de Wald III

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

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