Statistique Mathématique S6
Université de Strasbourg
L3 DUAS et L3 Mathématiques appliquées
Résumé du cours

Davide Giraudo

24 février 2025

1 Rappels sur les variables aléatoires discrètes

Nombre de façons de choisir p éléments dans un ensemble de n éléments

	avec ordre	sans ordre
avec répétition	n^p	C_{n+p-1}^p
sans répétition	A_n^p	C_n^p

Définition 1.1. *Soit $\mathcal{F}$ un ensemble de parties de Ω. On dit que $\mathcal{F}$ est une* **tribu** *si :*

1. *$\Omega \in \mathcal{F}$ (Ω est l'événement certain) ;*
2. *Si $A \in \mathcal{F}$, alors $A^c \in \mathcal{F}$;*
3. *Si $(A_n, n \in \mathbb{N})$ est une suite d'éléments de $\mathcal{F}$, alors $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{F}$.*

Définition 1.2. *Soient A et B deux événements tels que $\mathbb{P}(B) > 0$. La probabilité conditionnelle de A sachant B notée $\mathbb{P}(A|B)$ est définie par :*

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

Proposition 1.3. *Soient A et B deux événements tels que $\mathbb{P}(B) > 0$. On a*

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

Si de plus $\mathbb{P}(B^c) > 0$ (et donc $\mathbb{P}(B) < 1$), on a la formule de décomposition suivante :

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c),$$

et la formule de Bayes

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)}.$$

Définition 1.4. *On dit que deux événements A et B sont indépendants ssi*

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

Une famille au plus dénombrable d'événements $(A_i, i \in I)$, $(A_i \in \mathcal{F}, \forall i \in I)$ est indépendante ssi pour toute sous-famille finie $J \subseteq I$, on a

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

Dans cette section, $(\Omega, \mathcal{F}, \mathbb{P})$ désigne un espace de probabilité, tel que Ω soit au plus dénombrable.

Soit E un espace muni de la tribu $\mathcal{P}(E)$ (ensemble des sous-ensembles de E). On appelle **variable aléatoire discrète** à valeurs dans E toute application $X : \Omega \to E$. Le plus souvent $E = \mathbb{N}, \mathbb{Z}, \mathbb{R}$ ou $\mathbb{R}^d$.
On note P_X la fonction définie pour tout $A \in \mathcal{P}(E)$:

$$P_X(A) = \mathbb{P}(\{X \in A\}) = \mathbb{P}(X \in A).$$

P_X qui est une probabilité sur E, est appelée la **loi de la variable aléatoire** X ou encore sa **distribution**.
Soit $\{x_k, k = 1, 2, \dots\}$ l'ensemble des valeurs que X peut prendre. La loi de X est caractérisée par les données :

$$p_k = \mathbb{P}(\{X = x_k\}), \qquad k = 1, 2, \dots$$

Lorsque $E = \mathbb{R}$, on dit que X est une variable aléatoire réelle.
Propriété. *Soit g une fonction réelle telle que*

$$\sum_k |g(x_k)| p_k < \infty.$$

La variable $g \circ X$ que l'on note souvent $g(X)$ admet un moment d'ordre 1, avec

$$\mathbb{E}\left[g(X)\right] = \sum_k g(x_k) p_k.$$

- **Inégalité de Markov.** Soit $a > 0$. On a $\mathbb{P}(|X| \geqslant a) \leqslant \frac{\mathbb{E}[|X|]}{a}$.
- **Inégalité de Tchebychev.** Soit $a > 0$. On a $\mathbb{P}(|X| \geqslant a) \leqslant \frac{\mathbb{E}\left[X^2\right]}{a^2}$.
- **Inégalité de Jensen.** On suppose que X est intégrable. Soit ϕ une fonction convexe. Si $\mathbb{E}\left[\phi(X)\right]$ existe alors on a l'inégalité suivante :

$$\phi(\mathbb{E}\left[X\right]) \leqslant \mathbb{E}\left[\phi(X)\right].$$

- **Inégalité de Cauchy-Schwarz.** Soit (X, Y) un couple de variables réelles. On suppose que X^2 et Y^2 sont intégrables. Alors XY est intégrable et on a

$$\mathbb{E}\left[|XY|\right] \leqslant \sqrt{\mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right]}.$$

type de loi	paramètre(s)	valeurs prises	loi
constante		c	$\mathbb{P}(X=c)=1$
Bernoulli	$p\in[0,1]$	$\{0,1\}$	$\mathbb{P}(X=1)=p$ et $\mathbb{P}(X=0)=1-p$.
binomiale	$n\in\mathbb{N}^*, p\in[0,1]$	$\{0,\dots,n\}$	$\mathbb{P}(X=i)=C_n^i\,p^i\,(1-p)^{n-i}$
géométrique	$p\in]0,1[$	$\mathbb{N}^*$	$\mathbb{P}(X=n)=p(1-p)^{n-1}$
Poisson	$\lambda>0$	$\mathbb{N}$	$\mathbb{P}(X=n)=e^{-\lambda}\frac{\lambda^n}{n!}$

TABLE 1 – Lois discrètes usuelles

Définition 1.5. *La fonction génératrice de X est la fonction $G_X:[0,1]\to[0,1]$ définie par*

$$G_X(s)=\mathbb{E}\left[s^X\right]=\sum_{n=0}^{\infty}s^n\mathbb{P}(X=n),\quad s\in[0,1],$$

avec la convention $0^0:=1$.

Définition 1.6. *Soient X,Y deux variables discrètes définies sur le même espace probabilisé. La loi conditionnelle de Y sachant X notée $\mathcal{L}(Y|X)$ est la donnée de $\mathbb{P}(Y=y|X=x)$ pour x et y appartenant respectivement au support de la loi de X et de la loi de Y.*

2 Variables aléatoires continues

Définition 2.1. *On dit que f est une densité si f est positive, intégrable et d'intégrale égale à 1.*

Définition 2.2. *Soit X une variable réelle. On dit que X a pour densité f_X si pour tout borélien B,*

$$\mathbb{P}\left(X\in B\right)=\int_B f_X(x)dx$$

On dit alors que X est continue.

Théorème 2.3. *Si X admet la densité f_X et si g est une fonction sur $\mathbb{R}$ telle que*

$$\int_{\mathbb{R}}|g(x)|f_X(x)dx<\infty$$

alors $g(X)$ admet un moment d'ordre 1 avec

$$\mathbb{E}\left[g(X)\right]=\int_{\mathbb{R}}g(x)f_X(x)dx<\infty.$$

En particulier,

1. *si* $\int_{\mathbb{R}} |x| f_X(x)dx < \infty$, *alors* X *admet un moment d'ordre 1, avec*

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

2. *Si* I *est un intervalle, alors*

$$\mathbb{P}(X \in I) = \int_I f_X(x)dx < \infty.$$

Définition 2.4. *On appelle fonction de répartition de* X *(à valeurs réelles) la fonction* $F_X : \mathbb{R} \to [0,1]$ *donnée par :*

$$F_X(x) = P_X(]-\infty, x]) = \mathbb{P}(\{X \leqslant x\}), x \in \mathbb{R}.$$

Ainsi, la fonction F_X est croissante avec $\lim_{x\to-\infty} F_X(x) = 0 \quad$ et $\quad \lim_{x\to+\infty} F_X(x) = 1$ et F_X est continue à droite, sa limite à gauche au point x vaut

$$F_X(x-)\left(= \lim_{y\to x^-} F_X(y)\right) = \mathbb{P}(\{X < x\}) = P_X(]-\infty, x)).$$

Définition 2.5. *On dit que* $F\colon \mathbb{R} \to [0,1]$ *est une fonction de répartition si* F *est croissante avec* $\lim_{x\to-\infty} F(x) = 0 \quad$ *et* $\quad \lim_{x\to+\infty} F(x) = 1$ *et* F *est continue à droite, et a une limite à gauche en tout point* x*.*

Proposition 2.6. *La fonction de répartition* F_X *caractérise la loi de* X*.*

Proposition 2.7. *On suppose que* X *a une fonction de répartition* F_X *continue et dérivable* λ*-presque partout. Alors* X *admet pour densité* $f_X = F'_X$*.*

Nom de la loi	**paramètre(s)**	**densité**	**espérance**
Uniforme	$a < b$	$\frac{1}{b-a}\mathbf{1}_{]a,b[}(x)$	$\frac{a+b}{2}$
Exponentielle	$\theta > 0$	$f_X(x) = \theta e^{-\theta x}\mathbf{1}_{\mathbb{R}^+}(x)$	$\frac{1}{\theta}$
Gamma	$\lambda, \alpha > 0$	$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}\mathbf{1}_{\mathbb{R}_+}(x)$	$\frac{\alpha}{\lambda}$
Cauchy		$f_X(x) = \frac{1}{\pi}\frac{1}{1+x^2}$	
gaussienne	$\mu \in \mathbb{R}, \sigma^2 > 0$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ

TABLE 2 – Lois à densité usuelles

Proposition 2.8. *Soient* $X_1, \ldots, X_n$ *des variables réelles admettant toutes un moment d'ordre 1, et soient* $a_1, \ldots, a_n$ *des réels quelconques. Alors la variable* $\sum_{i=1}^n a_i X_i$ *admet également un moment d'ordre 1, et*

$$\mathbb{E}\left[\sum_{i=1}^n a_i X_i\right] = \sum_{i=1}^n a_i \mathbb{E}[X_i].$$

Proposition 2.9 (Inégalité de Markov)**.** *Soit X une variable réelle qui admet un moment d'ordre 1. Pour tout $a > 0$, on a*

$$\mathbb{P}(|X| \geqslant a) \leqslant \frac{\mathbb{E}\left[|X|\right]}{a}.$$

Définition 2.10. *On dit qu'une variable réelle X admet un moment d'ordre n si $\mathbb{E}\left[|X|^n\right] < \infty$.*

Proposition 2.11. *Soient $0 < p < q$. Si $\mathbb{E}\left[|X|^q\right] < \infty$, alors $\mathbb{E}\left[|X|^p\right] < \infty$. En particulier, si X admet un moment d'ordre n, alors elle admet des moments de tous ordres $m \leqslant n$.*

Définition 2.12. *Si X admet un moment d'ordre 2, alors elle admet un moment d'ordre 1. On pose alors*

$$\mathrm{Var}\left(X\right) := \mathbb{E}\left[X^2\right] - \left[\mathbb{E}\left[X\right]\right]^2,$$

et on appelle cette quantité la variance de X. On a en plus

$$\mathrm{Var}\left(X\right) = \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right].$$

Par conséquent $\mathrm{Var}\left(X\right) \geqslant 0$ et on appelle $\sigma_X := \sqrt{\mathrm{Var}\left(X\right)}$ l'écart-type de X.

Proposition 2.13.

1. *Une variable aléatoire est constante avec probabilité 1 ssi sa variance est nulle. La variable est alors égale à sa moyenne avec probabilité 1.*
2. *Soit X une variable réelle ayant un moment d'ordre 2, et soient a et b deux réels. Alors $aX + b$ admet aussi un moment d'ordre 2, avec*

$$\mathrm{Var}\left(aX + b\right) = a^2 \mathrm{Var}\left(X\right).$$

Proposition 2.14. *Soit X une variable qui admet un moment d'ordre 2. Pour tout réel $a > 0$, on a*

$$\mathbb{P}(|X - \mathbb{E}\left[X\right]| \geqslant a) \leqslant \frac{\mathrm{Var}\left(X\right)}{a^2}.$$

Proposition 2.15. *Soit $F : \mathbb{R} \to [0,1]$ la fonction de répartition associée à la loi de probabilité μ sur $\mathbb{R}$, i.e., $F(x) = \mu(]-\infty, x])$. Soit $F^{\leftarrow} : [0,1] \to \mathbb{R}$ l'inverse continue à gauche de F, i.e.,*

$$F^{\leftarrow}(x) = \inf\{y \in \mathbb{R} : F(y) > x\}, x \in]0,1[.$$

Si X suit une loi uniforme sur $[0,1]$, alors la variable $Y = F^{\leftarrow}(X)$ suit la loi μ.

3 Vecteurs aléatoires

Définition 3.1. *Soit $X = (X_1, \ldots, X_N)$ un vecteur aléatoire à valeurs dans $\mathbb{R}^N$. La fonction de répartition de X est*

$$\begin{aligned} F_X(x_1, \ldots, x_N) &= P_X(]-\infty, x_1] \times \cdots \times]-\infty, x_N]) \\ &= \mathbb{P}(X_1 \leqslant x_1, \ldots, X_N \leqslant x_N), \qquad (x_1, \ldots, x_N) \in \mathbb{R}^N. \end{aligned}$$

Définition 3.2. *On dit que $X = (X_1, \dots, X_N)$ a pour densité f_X si pour tous $x_1, \dots, x_N \in \mathbb{R}$,*

$$F_X(x_1, \dots, x_N) = \int \cdots \int_{]-\infty, x_1] \times \dots]-\infty, x_N]} f_X(u_1, \dots, u_N) du_1 \dots du_N.$$

Proposition 3.3. *Si $X = (X_1, \dots, X_N)$ a pour densité f_X, alors λ-presque partout,*

$$f_X(x_1, \dots, x_N) = \frac{\partial^N F}{\partial x_1 \dots \partial x_N} \mathbb{P}(X_1 \leqslant x_1, \dots, X_N \leqslant x_N), \qquad (x_1, \dots, x_N) \in \mathbb{R}^N.$$

Théorème 3.4. *Soit $X = (X_1, \dots, X_N)$ ayant pour densité f_X. Si une fonction $g : \mathbb{R}^N \to \mathbb{R}$ est telle que*

$$\int \cdots \int |g(x_1, \dots, x_n)| \, f_X(x_1, \dots, x_N) dx_1 \dots dx_N < \infty$$

alors la variable réelle $g(X_1, \dots, X_N)$ admet un moment d'ordre 1, avec

$$\mathbb{E}[g(X_1, \dots, X_N)] = \int \cdots \int_{\mathbb{R}^N} g(x_1, \dots, x_N) \, f_X(x_1, \dots, x_N) dx_1 \dots dx_N.$$

En particulier, pour tout borélien B de $\mathbb{R}^N$:

$$\mathbb{P}(X \in B) = \int \cdots \int_B f_X(x_1, \dots, x_N) dx_1 \dots dx_N.$$

Définition 3.5. *Soit $X = (X_1, \dots, X_N)$ un vecteur aléatoire. Si $X_1, \dots, X_N$ admettent toutes un moment d'ordre 1, alors le vecteur*

$$\mathbb{E}[X] = (\mathbb{E}[X_1], \dots, \mathbb{E}[X_N])$$

est appelé l'espérance de X.

Définition 3.6. *Soient X et Y deux variables réelles admettant toutes des moments d'ordre 2. Alors*

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

est appelée la covariance entre X et Y.

Proposition 3.7. *Si X et Y admettent un moment d'ordre 2, alors*

1. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
2. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
3. $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$
4. $\mathrm{Cov}(aX + b, cY + d) = a\,c\,\mathrm{Cov}(X, Y)$ *pour tous réels a, b, c, d,*
5. $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$.

Définition 3.8. *Si X et Y admettent des moments d'ordre 2, alors*

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

est appelé le coefficient de corrélation entre X et Y.

Définition 3.9. *Soit $X = (X_1, \dots X_N)$ un vecteur aléatoire telle que chaque composante admet un moment d'ordre 2. On appelle*

$$\Sigma = \begin{pmatrix} \mathrm{Var}\,(X_1) & \mathrm{Cov}(X_1, X_2) & \dots & \mathrm{Cov}(X_1, X_N) \\ \mathrm{Cov}(X_1, X_2) & \mathrm{Var}\,(X_2) & \dots & \mathrm{Cov}(X_2, X_N) \\ \vdots & & \dots. & \\ \vdots & & \dots. & \\ \vdots & & \dots. & \\ \mathrm{Cov}(X_1, X_N) & \mathrm{Cov}(X_2, X_N) & \dots & \mathrm{Var}\,(X_N) \end{pmatrix}$$

la matrice de variances-covariances de X. Il s'agit de la matrice symétrique $\Sigma = (\sigma_{ij})_{1 \leqslant i,j \leqslant N}$ avec $\sigma_{ij} = \mathrm{Cov}(X_i, X_j)$.

Proposition 3.10. *La matrice Σ est positive au sens que pour tout $a = (a_1, \dots, a_N)^\top \in \mathbb{R}^N$,*

$$a^\top \Sigma a = \sum_{1 \leqslant i,j \leqslant N} \sigma_{ij} a_i \, a_j \geqslant 0.$$

Proposition 3.11. *1. Si $X_1, \dots, X_N$ sont des variables réelles admettant toutes un moment d'ordre 2, alors*

$$\mathrm{Var}\,(X_1 + \dots + X_N) = \sum_{i=1}^{N} \mathrm{Var}\,(X_i) + 2 \sum_{1 \leqslant i < j \leqslant N} \mathrm{Cov}(X_i, X_j).$$

2. Si $X_1, \dots, X_M, Y_1, \dots, Y_N$ sont des variables réelles admettant toutes un moment d'ordre 2, alors pour tous réels $a_1, \dots, a_M, b_1, \dots, b_N$:

$$\mathrm{Cov}\left(\sum_{i=1}^{M} a_i X_i, \sum_{j=1}^{N} b_j Y_j \right) = \sum_{i=1}^{M} \sum_{j=1}^{N} a_i \, b_j \, \mathrm{Cov}\,(X_i, Y_j)\,.$$

Théorème 3.12. *Formule du changement de variables Soit X un vecteur aléatoire à valeurs dans un ouvert $\Delta \subset \mathbb{R}^N$ (souvent $\Delta = \mathbb{R}^N$) qui admet comme densité $f_X = f_X \mathbf{1}_\Delta$. Soit*

$$h : \Delta \to D$$

(avec $D \subset \mathbb{R}^N$) une fonction bijective, $\mathcal{C}^1$, à réciproque $\mathcal{C}^1$ (i.e., les dérivées partielles existent et sont continues). On note $h^{\leftarrow}$ la réciproque de h (ou h^{-1}). Alors le vecteur aléatoire $Y = h \circ X = h(X)$ a pour densité

$$f_Y(y) = \frac{f_X(h^{\leftarrow}(y))}{|\det\left(\mathrm{Jac}\,(h)\,(h^{\leftarrow}(y))\right)|}$$

où $\mathrm{Jac}\,(h)\,(x)$ *est le jacobien de* h *en* x *: en notant* $h = (h_1, \ldots, h_N)$,

$$\mathrm{Jac}\,(h)\,(x) = \begin{pmatrix} \frac{\partial h_1(x_1,\ldots,x_N)}{\partial x_1} & \cdots & \frac{\partial h_1(x_1,\ldots,x_N)}{\partial x_N} \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ \frac{\partial h_N(x_1,\ldots,x_N)}{\partial x_1} & \cdots & \frac{\partial h_N(x_1,\ldots,x_N)}{\partial x_N} \end{pmatrix}$$

Définition 3.13. *Soit* $X = (X_1, \ldots, X_N)$ *un vecteur aléatoire. On appelle loi marginale la loi de tout sous-vecteur* $(X_{i_1}, \ldots, X_{i_k})$ *(avec* $k < N$*) extrait de* X*.*

Proposition 3.14. *Soit*

$$\begin{aligned} F_{X_1,\ldots,X_N}(x_1, \ldots, x_N) &= P_{(X_1,\ldots,X_N)}(]-\infty, x_1] \times \cdots \times]-\infty, x_N]) \\ &= \mathbb{P}(X_1 \leqslant x_1, \ldots, X_N \leqslant x_N) \end{aligned}$$

la fonction de répartition de $(X_1, \ldots, X_N)$*. Soit* $1 \leqslant k < N$*. Alors la fonction de répartition de* $(X_1, \ldots, X_k)$ *est*

$$F_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \lim_{x_{k+1}\to\infty,\ldots,x_N\to\infty} F_{X_1,\ldots,X_N}(x_1, \ldots, x_N).$$

On écrit souvent

$$F_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = F_{X_1,\ldots,X_N}(x_1, \ldots, x_k, +\infty, \ldots, +\infty).$$

En particulier, la loi (marginale) de X_i *est donnée par sa fonction de répartition*

$$F_{X_i}(x_i) = F_{X_1,\ldots,X_N}(+\infty, \ldots, +\infty, x_i, +\infty, \ldots, +\infty).$$

Proposition 3.15. *Si* $f_{(X_1,\ldots,X_n)}$ *est la densité du vecteur aléatoire* $X = (X_1, \ldots, X_n)$*, alors le vecteur aléatoire* $(X_1, \ldots, X_k)$ *admet la densité*

$$f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \int \cdots \int_{\mathbb{R}^{n-k}} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) dx_{k+1} \ldots dx_n.$$

Définition 3.16. *Si* $f_X(x) \neq 0$ *alors la fonction de la variable* y *:*

$$f_{Y|\{X=x\}}(y) = \frac{f_{(X,Y)}(x, y)}{f_X(x)}$$

est une densité. Elle est appelée la densité de la loi conditionnelle de Y *sachant* $X = x$*.*

On a

$$\mathbb{P}(Y \in A | X = x) = \int_A f_{Y|\{X=x\}}(y) dy.$$

ainsi que la formule de Bayes :

$$f_{X|\{Y=y\}}(x) = \frac{f_{(X,Y)}(x, y)}{f_Y(y)} = \frac{f_{Y|\{X=x\}}(y) f_X(x)}{\int f_{Y|\{X=u\}}(y) f_X(u) du}.$$

Définition 3.17. *Soit ϕ une fonction. On suppose que $\phi(X,Y)$ est intégrable et que X est une variable continue. Définissons la fonction ψ par $\psi(x)=0$ si $f_X(x)=0$ et*

$$\psi(x)=\int \phi(x,y)f_{Y|\{X=x\}}(y)dy \quad si \quad f_X(x)\neq 0.$$

La variable $\psi(X)$ est l'espérance conditionnelle de $\phi(X,Y)$ sachant X. On la note $\mathbb{E}\left[\phi(X,Y)|X\right]$. Par convention, on note $\psi(x)=\mathbb{E}\left[\phi(X,Y)|X=x\right]$.

Définition 3.18. *Soient X et Y deux variables aléatoires à valeurs dans $\mathbb{R}^d$ et $\mathbb{R}^{d'}$, respectivement. On dit que X et Y sont indépendantes si pour tout $A\in\mathcal{B}\left(\mathbb{R}^d\right)$ et tout $B\in\mathcal{B}\left(\mathbb{R}^{d'}\right)$,*

$$\mathbb{P}(X\in A, Y\in B)=\mathbb{P}(X\in A)\mathbb{P}(Y\in B).$$

Quand on exprime l'indépendance en termes des distributions, on obtient que deux variables X et Y sont indépendantes ssi $P_{(X,Y)}-$ la loi du couple $(X,Y)-$ vérifie

$$P_{(X,Y)}(A\times B)=P_X(A)P_Y(B), \forall A\in\mathcal{B}\left(\mathbb{R}^d\right), B\in\mathcal{B}\left(\mathbb{R}^{d'}\right).$$

Autrement dit, la loi $P_{(X,Y)}$ sur $\mathcal{B}\left(\mathbb{R}^d\times\mathbb{R}^{d'}\right)$ est la loi produit $P_X\otimes P_Y$.

Proposition 3.19. *Pour que deux variables aléatoires X et Y soient indépendantes, il faut et il suffit que*

$$\mathbb{E}\left[g_1(X)g_2(Y)\right]=\mathbb{E}\left[g_1(X)\right]\mathbb{E}\left[g_2(Y)\right],$$

pour toute fonction bornée $g_1\colon\mathbb{R}^d\to\mathbb{R}$ et toute fonction bornée $g_2\colon\mathbb{R}^{d'}\to\mathbb{R}$.

Définition 3.20. *On dit que n variables $X_1,\dots,X_n$ sont indépendantes si*

$$\mathbb{P}(X_1\in A_1,\dots,X_n\in A_n)=\mathbb{P}(X_1\in A_1)\dots\mathbb{P}(X_n\in A_n),\quad \forall A_1\in\mathcal{E}_1,\dots,\forall A_n\in\mathcal{E}_n.$$

Ceci équivaut à

$$\mathbb{E}\left[g_1(X_1)\dots g_n(X_n)\right]=\mathbb{E}\left[g_1(X_1)\right]\dots\mathbb{E}\left[g_n(X_n)\right]$$

pour toutes les fonctions bornées $g_k\colon E_k\to\mathbb{R}$.

Théorème 3.21. *Les variables réelles $X_1,\dots,X_n$ sont indépendantes ssi*

$$F_{(X_1,\dots,X_n)}(x_1,\dots,x_n)=F_{X_1}(x_1)\dots F_{X_n}(x_n)\qquad\forall(x_1,\dots,x_n)\in\mathbb{R}^n.$$

Théorème 3.22. *Soit $(X_1,\dots,X_n)$ un vecteur aléatoire à densité. Alors $X_1,\dots,X_n$ sont indépendantes ssi*

$$f_{(X_1,\dots,X_n)}(x_1,\dots,x_n)=f_{X_1}(x_1)\dots f_{X_n}(x_n)\qquad\forall(x_1,\dots,x_n)\in\mathbb{R}^n$$

pour presque tout $(x_1,\dots,x_n)\in\mathbb{R}^n$.

Proposition 3.23. *Si la densité du vecteur aléatoire $(X_1,\dots,X_n)$ se factorise*

$$f_{(X_1,\dots,X_n)}(x_1,\dots,x_n)=g_{X_1}(x_1)\dots g_{X_n}(x_n)$$

alors $X_1,\dots,X_n$ sont indépendantes.

Soit X une variable de densité f_X à valeurs dans $\mathbb{R}^N$ muni du produit scalaire $\langle t, X\rangle = \sum_{j=1}^N t_j X_j$.

Définition 3.24. *L'application $\phi_X : \mathbb{R}^N \to \mathbb{C}$ donnée par*

$$\phi_X(t) = \mathbb{E}\left[e^{i\langle t,X\rangle}\right] = \int_{\mathbb{R}^N} e^{i\langle t,x\rangle} f_X(x)dx$$

s'appelle la fonction caractéristique de X.

Théorème 3.25. *Deux variables aléatoires X et Y ont même loi ssi $\phi_X = \phi_Y$. Autrement dit, comme son nom l'indique, la fonction caractéristique ϕ_X caractérise la loi de X.*

Théorème 3.26. *Les variables réelles $X_1, \ldots, X_n$ sont indépendantes ssi*

$$\forall t = (t_1, \ldots, t_n) \in \mathbb{R}^n, \quad \phi_{(X_1,\ldots,X_n)}(t) = \prod_{k=1}^n \phi_{X_k}(t_k).$$

Proposition 3.27. *Supposons que la variable réelle X admet un moment d'ordre $n \geqslant 1$. Alors ϕ_X est de classe $\mathcal{C}^n$ (n fois dérivable et la dérivée n-ème continue) et*

$$\phi_X^{(n)}(t) = i^n \mathbb{E}\left[X^n e^{itX}\right], \qquad t \in \mathbb{R}.$$

En particulier

$$\mathbb{E}\left[X^n\right] = \frac{\phi_X^{(n)}(0)}{i^n}.$$

4 Théorèmes fondamentaux en probabilité

Définition 4.1. *On dit que la suite $(X_n)_{n\geqslant 1}$ converge en probabilité vers une variable X si pour tout $\varepsilon > 0$*

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

Proposition 4.2. *Si $X_n \to X$ en probabilité et si f est une fonction continue sur $\mathbb{R}$, alors $f(X_n) \to f(X)$ en probabilité.*

Définition 4.3. *Pour tout $p \geqslant 1$, on dit qu'une suite $(X_n)_{n\geqslant 1}$ de variables converge en moyenne d'ordre p vers une variable X si*

$$\lim_{n\to\infty} \mathbb{E}\left[|X_n - X|^p\right] = 0.$$

Lorsque $p = 2$ on dit également que X_n converge vers X en moyenne quadratique.

Proposition 4.4. *Si $X_n \to X$ en moyenne d'ordre p, alors la convergence a encore lieu en probabilité.*

Définition 4.5. *La suite $(X_n)_{n\geqslant 1}$ converge presque sûrement vers X si $\mathbb{P}\left(\{\lim_{n\to\infty} X_n = X\}\right) = 1$.*

Si pour tout $\varepsilon > 0$

$$\sum_{n=1}^{+\infty} \mathbb{P}(|X_n - X| > \varepsilon) < \infty$$

alors $(X_n)_{n\geqslant 1}$ converge presque sûrement vers X.

Loi faible des grands nombres. *Soit $(X_n)_{n\geqslant 1}$ une suite de variables indépendantes, intégrables et de même loi (continue ou discrète). On note $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ la moyenne empirique. Alors la moyenne empirique converge en probabilité vers $\mathbb{E}[X_1]$, i.e., pour tout $\varepsilon > 0$:*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - \mathbb{E}[X_1]\right| > \varepsilon\right) = 0.$$

Loi forte des grands nombres. *Soit $(X_n)_{n\geqslant 1}$ une suite de variables indépendantes, intégrables et de même loi (continue ou discrète). On note $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ la moyenne empirique. Alors la moyenne empirique converge presque sûrement vers $\mathbb{E}[X_1]$.*

5 Convergence en loi

Définition 5.1. *On dit qu'une suite de variables $(X_i)_{i\geqslant 1}$ à valeurs dans $\mathbb{R}^N$ (donc une suite de vecteurs aléatoires) converge en loi vers une variable X (à valeurs dans $\mathbb{R}^N$) si, pour toute fonction f continue bornée sur $\mathbb{R}^N$,*

$$\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)].$$

Proposition 5.2. *Si $(X_n)_{n\geqslant 1}$ est une suite de variables réelles qui converge en probabilité vers une variable X alors $(X_n)_{n\geqslant 1}$ converge également vers X en loi.*

Théorème 5.3. *Soient $(X_n)_{n\geqslant 1}$ une suite de variables réelles. Soit F_X la fonction de répartition d'une variable réelle X. Les assertions suivantes sont équivalentes :*

1. *$(X_n)_{n\geqslant 1}$ converge en loi vers X,*
2. *si x est un point de continuité de F_X, alors $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$.*

Théorème 5.4. *Soit $(X_i)_{i\geqslant 1}$ une suite de variables réelles indépendantes et identiquement distribuées avec $\mathbb{E}[X^2] < \infty$. On note $S_n = X_1 + \cdots + X_n$ la somme partielle au rang n et $\sigma^2 = \mathrm{Var}(X)$. Alors*

$$\frac{S_n - n\mathbb{E}[X]}{\sqrt{n}}$$

converge en loi vers une $\mathcal{N}(0, \sigma^2)$.

6 Statistiques exhaustives et modèles dominés

Définition 6.1. *On appelle modèle statistique tout couple $\left(\mathbb{X}, (\mathbb{P}_\theta)_{\theta\in\theta}\right)$ où*

- *$\mathbb{X}$ est un ensemble dit espace des observations,*

- Θ *est un ensemble dit espace des paramètres et* $(\mathbb{P}_\theta)_{\theta\in\Theta}$ *est une famille de probabilités définies sur une tribu fixée de parties de* $\mathbb{X}$.

Définition 6.2. *Une mesure* μ *sur* $(\mathbb{X},\mathcal{A})$ *est* $\sigma-$*finie ssi il existe une suite* $\{A_n\}_{n\geqslant 1}$ *d'événements de* $\mathcal{A}$ *telle que* $\cup_{n\geqslant 1}A_n=\mathbb{X}$ *et* $\forall n\geqslant 1$, $\mu(A_n)<\infty$, *autrement dit,* $\mathbb{X}$ *est une union dénombrable d'événements de mesure finie.*

Définition 6.3. *Un modèle* $(\mathbb{X},(\mathbb{P}_\theta)_{\theta\in\Theta})$ *est dit dominé s'il existe une mesure* σ*-finie* μ *vérifiant*

$$\forall\theta\in\Theta:\mathbb{P}_\theta\ll\mu,$$

i.e., $\forall A$ *mesurable* $\mu(A)=0$ *implique que pour tout* $\theta\in\Theta$, $\mathbb{P}_\theta(A)=0$.

Dans le cadre de la statistique inférentielle, on dispose d'une observation x régie par une loi de probabilité $\mathbb{P}_\theta$ dépendant d'un paramètre inconnu θ et que le principe de base adopté consiste à n'effectuer d'inférence sur la valeur θ qu'au travers de l'information contenue dans x.

Exemple 6.4. Supposons que l'observation x soit un échantillon $(x_1,\ldots,x_n)$ de la loi de Bernoulli de paramètre θ. Ainsi l'espace des observations $\mathbb{X}$ est l'ensemble $\{0,1\}^n$ et l'espace des paramètres Θ est l'intervalle $[0,1]$. Intuitivement, comme il s'agit d'une situation d'échantillonnage, il semble que l'on ne perde rien, vis-à-vis de θ, à considérer le nombre de composantes x_i égales à 1 à la place de l'échantillon complet $(x_1,\ldots,x_n)$. Notons ϕ la statistique correspondante, soit $\phi:(x_1,\ldots,x_n)\mapsto\sum_{i=1}^n x_i$. $\forall x=(x_1,\ldots,x_n)\in\mathbb{X}$ et $\forall y\in\{0,1,\ldots,n\}$, on a

$$\begin{aligned}\mathbb{P}_\theta[\{x\}|\phi=y] &= \frac{\mathbb{P}_\theta[\{x\}\cap\phi=y]}{\mathbb{P}_\theta[\phi=y]}\\ &= \frac{\prod_{i=1}^n\theta^{x_i}(1-\theta)^{1-x_i}}{C_n^y\theta^y(1-\theta)^{n-y}} \quad \text{si } \sum x_i=y \text{ (et 0 sinon)}\\ &= \frac{1}{C_n^y} \quad \text{si } \sum x_i=y \text{ (et 0 sinon).}\end{aligned}$$

Clairement cette probabilité conditionnelle est indépendante de θ ainsi la statistique ϕ est bien exhaustive.

Définition 6.5. *On dit qu'une statistique* ϕ *définie sur un modèle statistique* $(\mathbb{X},(\mathbb{P}_\theta)_{\theta\in\Theta})$ *est exhaustive s'il existe une version de la probabilité conditionnelle* $\mathbb{P}(.|\phi)$ *commune à chacune des probabilités* $\mathbb{P}_\theta$, *i.e., indépendante de* θ. *Si tel est le cas,* $\mathcal{A}$ *étant la tribu considérée sur* $\mathbb{X}$ *et* $(\mathbb{Y},\mathcal{B})$ *étant l'espace image de* ϕ, *on a*

$$\forall\theta\in\Theta,\forall A\in\mathcal{A},\forall B\in\mathcal{B}:\mathbb{P}_\theta(A\cap\phi^{-1}(B))=\int_B\mathbb{P}(A|\phi=y)\phi(\mathbb{P}_\theta)(dy).$$

Définition 6.6 (Vraisemblance)**.** *La vraisemblance d'un échantillon* $(x_1,\ldots,x_n)$ *est*

$$\mathcal{L}(\theta,x_1,\ldots,x_n)=\begin{cases}\mathbb{P}(X_1=x_1,\ldots,X_n=x_n,\theta) & \text{si les } X_i \text{ sont de lois discrètes.}\\ f_{X_1,\ldots,X_n}(x_1,\ldots,x_n,\theta) & \text{si les } X_i \text{ sont de lois continues.}\end{cases}$$

Dans le cadre d'un modèle d'échantillon indépendant et identiquement distribué, cela devient

$$\mathcal{L}(\theta, x_1, \ldots, x_n) = \begin{cases} \prod_{i=1}^{n} \mathbb{P}(X = x_i, \theta) & \text{si les } X_i \text{ sont de lois discrètes.} \\ \prod_{i=1}^{n} f(x_i, \theta) & \text{si les } X_i \text{ sont de lois continues.} \end{cases}$$

Les densités et les probabilités dépendent du paramètre θ.

Théorème 6.7 (Neyman-Fisher)**.** *Considérons un modèle statistique $(\mathbb{X}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ dominé par une mesure $\sigma-$finie μ et de vraisemblance f. Toute statistique ϕ à valeurs dans un espace $\mathbb{Y}$ est exhaustive ssi il existe une application g de $\mathbb{Y} \times \Theta$ dans $\mathbb{R}_+$ et une application h de $\mathbb{X}$ dans $\mathbb{R}_+$ telles que la vraisemblance f s'écrive*

$$\forall x \in \mathbb{X}, \forall \theta \in \Theta : f(x, \theta) = g(\phi(x), \theta).h(x).$$

Définition 6.8. *Soit $(\mathbb{X}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ un modèle statistique. Une statistique exhaustive ϕ est dite exhaustive minimale si, pour toute statistique exhaustive ϕ', la statistique ϕ est une fonction de ϕ' (autrement dit il existe une fonction ξ telle que $\phi = \xi \circ \phi'$).*

Théorème 6.9 (Lehmann et Scheffé (1950))**.** *Supposons que le modèle statistique $(\mathbb{X}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ admette une vraisemblance f. Soit ϕ une statistique sur ce modèle. Si on a équivalence entre*

$$\phi(x) = \phi(y) \Longleftrightarrow \theta \longmapsto \frac{f(x, \theta)}{f(y, \theta)} \quad \text{est une fonction indépendante de } \theta$$

alors ϕ est une statistique exhaustive minimale pour θ.

Définition 6.10. *Soit $(\mathbb{X}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ un modèle statistique. Une statistique sur ce modèle est dite libre si sa loi ne dépend pas de θ.*

Définition 6.11. *Soit $(\mathbb{X}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ un modèle statistique et soit F un sous-espace vectoriel de l'espace des fonctions $(\mathbb{P}_\theta)_{\theta\in\Theta}-$ intégrables de $\mathbb{X}$ dans $\mathbb{R}$. On dit qu'une statistique ϕ $(\mathbb{X} \longrightarrow \mathbb{Y})$ est $F-$complète si, pour toute fonction g définie sur $\mathbb{Y}$ telle que $g \circ \phi \in F$, on a l'implication*

$$\forall \theta \in \Theta : \int_{\mathbb{X}} g \circ \phi d\mathbb{P}_\theta = 0 \Longrightarrow \forall \theta \in \Theta : g \circ \phi = 0 \quad \mathbb{P}_\theta - ps$$

soit encore

$$\forall \theta \in \Theta : \int_{\mathbb{Y}} g d\phi(\mathbb{P}_\theta) = 0 \Longrightarrow \forall \theta \in \Theta : g = 0 \quad \phi(\mathbb{P}_\theta) - ps.$$

Dans le cas où F est l'espace vectoriel des fonctions bornées, on parle de statistique bornée complète ou quasi-complète et, dans le cas où F consiste en tout l'espace des fonctions $(\mathbb{P}_\theta)_{\theta\in\Theta}-$ intégrables, on parle de statistique complète.

Enfin, lorsque ϕ est l'identité sur $\mathbb{X}$, on dit que le modèle $(\mathbb{X}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ est $F-$complet (respectivement quasi-complet ou complet). Dans ce cas, pour toute fonction g de F (respectivement bornée ou $(\mathbb{P}_\theta)_{\theta\in\Theta}-$intégrable), on a l'implication

$$\forall \theta \in \Theta : \int_{\mathbb{X}} g d\mathbb{P}_\theta = 0 \Longrightarrow \forall \theta \in \Theta : g = 0 \quad \mathbb{P}_\theta - ps.$$

Théorème 6.12. *Soit $(\mathbb{X}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ un modèle exponentiel de dimension n et de vraisemblance L de la forme*

$$\forall x \in \mathbb{X}, \forall \theta \in \Theta : L(x,\theta) = \beta(\theta).\xi(x).\exp(\phi(x).\alpha(\theta))$$

par rapport à une mesure dominante μ. Si la partie $\alpha(\Theta)$ de $\mathbb{R}^n$ est d'intérieur non vide, alors la statistique ϕ est complète.

Théorème 6.13. *Soit $(\mathbb{X}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ un modèle statistique. Si ϕ est une statistique exhaustive quasi-complète à valeurs dans $\mathbb{R}^k$, alors ϕ est une statistique exhaustive minimale.*

Théorème 6.14 (Basu)**.** *Si ϕ $(\mathbb{X} \longrightarrow \mathbb{Y})$ est une statistique exhaustive quasi-complète sur un modèle statistique $(\mathbb{X}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ alors ϕ est une statistique $(\mathbb{P}_\theta)_{\theta\in\Theta}-$ indépendante de toute statistique libre sur ce modèle.*

7 Estimation ponctuelle

Définition 7.1 (Statistique)**.** *Une statistique t est une fonction des observations $x_1, \ldots, x_n$: $t : \mathbb{R}^n \mapsto \mathbb{R}^d$ associant $t(x_1, \ldots, x_n)$ au point $(x_1, \ldots, x_n)$.*

Définition 7.2 (Estimateur)**.** *Un estimateur d'une grandeur θ est une statistique T_n à valeurs dans l'ensemble des valeurs possibles de θ. Une estimation de θ est une réalisation t_n de T_n.*

Définition 7.3 (Moments centrés et ordinaires)**.** *Les moments centrés et ordinaires d'une variable aléatoire X sont définis par*

$$\mu_k = \mathbb{E}\left[(X - \mathbb{E}[X])^k\right] \text{ et } m_k = \mathbb{E}\left[X^k\right].$$

Leur version empirique est

$$\widehat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^k \quad \text{et} \quad \widehat{m}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

Définition 7.4 (Moment Matching Estimation)**.** *La méthode des moments consiste à égaler les d premiers moments théoriques et leurs versions empiriques où d est la dimension du paramètre.*

Définition 7.5 (Maximum Likelihood Estimation)**.** *La méthode par maximum de vraisemblance consiste à trouver la valeur du paramètre maximisant la vraisemblance ou la log-vraisemblance, c'est à dire*

$$\arg\max_\theta \mathcal{L}(\theta, x_1, \ldots, x_n) \text{ ou } \arg\max_\theta \log\mathcal{L}(\theta, x_1, \ldots, x_n).$$

Des formules closes peuvent exister si les équations de vraisemblance $\frac{\partial\mathcal{L}}{\partial\theta}(\theta, .) = 0$ ou $\frac{\partial\log\mathcal{L}}{\partial\theta}(\theta, .) = 0$ possèdent des solutions explicites, sinon on a recours à une optimisation numérique.

8 Qualité d'un estimateur

Définition 8.1 (Biais). *Pour un estimateur T_n de θ, le biais se définit comme $b(T_n) = \mathbb{E}[T_n] - \theta$. T_n est dit sans biais si le biais vaut zéro, sinon il est dit biaisé. T_n est dit asymptotiquement sans biais si $\mathbb{E}[T_n] \to \theta$ lorsque $n \to +\infty$.*

Définition 8.2 (Erreur quadratique moyenne). *Pour un estimateur T_n de θ, l'erreur quadratique moyenne ("mean squared error") se définit comme $\mathrm{MSE}(T_n) = \mathbb{E}\left[(T_n - \theta)^2\right]$. On peut montrer que $\mathrm{MSE}(T_n) = \mathrm{Var}(T_n)) + (\mathbb{E}[T_n] - \theta)^2$.*

Définition 8.3 (Convergence en moyenne quadratique). *L'estimateur T_n converge en moyenne quadratique vers θ si et seulement si son erreur quadratique moyenne tend vers 0 quand n tend vers l'infini :*

$$\mathbb{E}\left[(T_n - \theta)^2\right] \to 0.$$

Définition 8.4 (Modèle paramétrique). *Un modèle paramétrique est un triplet $(\mathbb{X}, \mathcal{A}, \mathcal{P})^n$ où $\mathbb{X}$ est l'ensemble de valeurs possibles des observations (de la variable aléatoire X), $\mathcal{A}$ est la tribu des éléments observables, $\mathcal{P}$ l'ensemble des lois possibles pour X du type $\mathcal{P} = \{P_\theta, \theta \in \Theta \subset \mathbb{R}^d\}$. La puissance n souligne l'hypothèse indépendant et identiquement distribué.*

Le paramètre θ est inconnu mais on sait qu'il est à valeurs dans Θ. En pratique, $\mathcal{A}$ sera l'ensemble des parties de $\mathbb{X}$ lorsque $\mathbb{X}$ est discret ou l'ensemble des boréliens sinon. $\mathcal{A}$ ne sera pas mentionné.

Définition 8.5 (Hypothèses de modèles réguliers). *Les hypothèses d'un modèle régulier $(\mathbb{X}, \{P_\theta, \theta \in \Theta \subset \mathbb{R}^d\})$ pour une mesure μ sont les suivantes*

(H1) $\mathbb{X}$ ne dépend pas du paramètre θ, i.e., le support de P_θ est indépendant de θ.

(H2) la vraisemblance $\mathcal{L}$ associée à P_θ est positive, i.e., $\forall (x, \theta) \in \mathbb{X} \times \Theta, \mathcal{L}(x, \theta) > 0$.

(H3) $\ln L$ est deux fois dérivables par rapport à chaque composante θ_j de θ.

(H4) On peut inverser dérivée et somme de la manière suivante, pour toute fonction g mesurable

$$\frac{\partial}{\partial \theta_j} \int_A g(.)\mathcal{L}(., \theta) d\mu = \int_A g(.) \frac{\partial \mathcal{L}}{\partial \theta_j}(., \theta) d\mu \text{ et } \frac{\partial^2}{\partial \theta_j \partial \theta_k} \int_A g(.)\mathcal{L}(., \theta) d\mu = \int_A g(.) \frac{\partial^2 \mathcal{L}}{\partial \theta_j \partial \theta_k}(., \theta) d\mu.$$

Définition 8.6 (Score). *Sous (H1), (H2), (H3), le score pour la variable générique X est défini comme le vecteur gradient de $\ln L(X, \theta)$ (i.e., la dérivée de $\ln L(X, \theta)$ par rapport à θ) :*

$$Z_\theta = \nabla_\theta \ln L(X, \theta) \in \mathbb{R}^d.$$

Notons que l'estimateur de maximum de vraisemblance annule le score.

Définition 8.7 (Quantité d'information de Fisher – cas multidimensionnel). *Pour un échantillon $X_1, \ldots, X_n$ et sous (H1), (H2), (H3), la matrice d'information de Fisher $I_n(\theta)$ est donnée par*

$$I_n(\theta) = (\mathrm{Cov}(Z_{\theta,i}, Z_{\theta,j}))_{ij} = \left(\mathrm{Cov}\left(\frac{\partial \ln L}{\partial \theta_i}(\theta, X_1, \ldots, X_n), \frac{\partial \ln L}{\partial \theta_j}(\theta, X_1, \ldots, X_n)\right)\right)_{ij}.$$

Définition 8.8 (Quantité d'information de Fisher – cas réel $d = 1$)**.** *Pour $\theta \in \mathbb{R}$ et un échantillon $X_1, \dots, X_n$ et sous (H1), (H2), (H3), la quantité d'information de Fisher est donnée par*

$$I_n(\theta) = \mathrm{Var}\left(Z_\theta\right) = \mathrm{Var}\left(\frac{\partial}{\partial\theta}\log\mathcal{L}(\theta, X_1, \dots, X_n)\right).$$

Proposition 8.9 (Simplification – cas multidimensionnel)**.** *Sous (H4), la quantité d'information peut se réécrire*

$$I_n(\theta) = \left(-\mathbb{E}\left[\frac{\partial^2 \ln L}{\partial\theta_i\partial\theta_j}(\theta, X_1, \dots, X_n)\right]\right)_{ij}.$$

De plus l'information de Fisher est additive, ainsi

$$I_n(\theta) = nI_1(\theta) = \left(-n\mathbb{E}\left[\frac{\partial^2 \ln L}{\partial\theta_i\partial\theta_j}(\theta, X_1)\right]\right)_{ij}.$$

Proposition 8.10 (Simplification – cas réel $d = 1$)**.** *Sous (H4), la quantité d'information peut se réécrire*

$$I_n(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\mathcal{L}(\theta, X_1, \dots, X_n)\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log\mathcal{L}(\theta, X_1, \dots, X_n)\right].$$

De plus l'information de Fisher est additive, ainsi

$$I_n(\theta) = nI_1(\theta) = -n\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log\mathcal{L}(\theta, X)\right].$$

Proposition 8.11 (Inégalité de Fréchet-Darmois-Cramer-Rao – cas multidimensionnel)**.** *Sous (H1), (H2), (H3), (H4), pour tout estimateur $T_n \colon \mathbb{R}^n \mapsto \mathbb{R}^d$ de θ on a*

$$\mathrm{Var}\left(T_{n,i}\right) \geqslant \sum_{j=1}^{d}\sum_{k=1}^{d}(I_n(\theta)^{-1})_{jk}\frac{\partial\mathbb{E}\left[T_{n,i}\right]}{\partial\theta_j}\frac{\partial\mathbb{E}\left[T_{n,i}\right]}{\partial\theta_k}$$

Dans le cas d'un estimateur sans biais, cela se réduit à

$$\mathrm{Var}\left(T_{n,i}\right) \geqslant (I_n(\theta)^{-1})_{ii},$$

où $(I_n(\theta)^{-1})_{ii}$ est appelée de borne de Cramer-Rao.

Proposition 8.12 (Inégalité de Fréchet-Darmois-Cramer-Rao – cas réel $d = 1$)**.** *Si la loi des observations vérifie des conditions de régularité, alors pour tout estimateur T_n de θ on a*

$$\mathrm{Var}\left(T_n\right) \geqslant \frac{\left(\frac{\partial\mathbb{E}[T_n]}{\partial\theta}\right)^2}{I_n(\theta)}.$$

Dans le cas d'un estimateur sans biais, cela se réduit à

$$\mathrm{Var}\left(T_n\right) \geqslant \frac{1}{I_n(\theta)},$$

où $1/I_n(\theta)$ est appelée de borne de Cramer-Rao.

Définition 8.13 (Efficacité – cas multidimensionnel). *L'efficacité d'un estimateur T_n de θ est définie par*

$$\mathrm{Eff}\,(T_{n,i}) = \sum_{j=1}^{d}\sum_{k=1}^{d}(I_n(\theta)^{-1})_{jk}\frac{\partial \mathbb{E}\,[T_{n,i}]}{\partial \theta_j}\frac{\partial \mathbb{E}\,[T_{n,i}]}{\partial \theta_k}\frac{1}{\mathrm{Var}\,(T_{n,i})}.$$

T_n est dit efficace (resp. asymptotiquement efficace) si $\mathrm{Eff}\,(T_{n,i}) = 1$ *pour tout i (resp. si* $\mathrm{Eff}\,(T_{n,i}) \to 1$*).*

Définition 8.14 (Efficacité – cas réel $d = 1$). *L'efficacité d'un estimateur T_n de θ est définie par*

$$\mathrm{Eff}\,(T_n) = \frac{\left(\frac{\partial \mathbb{E}[T_n]}{\partial \theta}\right)^2}{I_n(\theta)\,\mathrm{Var}\,(T_n)}.$$

T_n est dit efficace (resp. asymptotiquement efficace) si $\mathrm{Eff}\,(T_n) = 1$ *(resp. si* $\mathrm{Eff}\,(T_n) \to 1$*).*

- Si T_n est un estimateur sans biais de θ, $\mathrm{Eff}\,(T_n) = \frac{1}{I_n(\theta)\,\mathrm{Var}(T_n)}$.
- Si un estimateur sans biais est efficace, sa variance est égale à la borne de Cramer-Rao, donc c'est forcément l'ESBVM (un ESBVM est presque sûrement unique).
- Il est possible qu'il n'existe pas d'estimateur efficace de θ. Alors, s'il existe un ESBVM de θ, sa variance est strictement supérieure à la borne de Cramer-Rao.

Théorème 8.15 (Loi forte des Grands Nombres). *Soit $(X_i)_{i\geqslant 1}$ une suite de vecteurs aléatoires indépendants, de même loi, admettant un vecteur moyenne m fini de dimension d. On note*

$$X_i = \begin{pmatrix} X_{i,1} \\ \vdots \\ X_{i,d} \end{pmatrix}\ ,\ \frac{1}{n}\sum_{i=1}^{n} X_i = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} X_{i,1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} X_{i,d} \end{pmatrix} \text{ et } m = \begin{pmatrix} m_1 \\ \vdots \\ m_d \end{pmatrix}.$$

On a alors

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to m \quad \text{p.s. dans } \mathbb{R}^d$$

Théorème 8.16 (Théorème limite central). *Soit $(X_i)_{i\geqslant 1}$ une suite de vecteurs aléatoires de dimension d indépendants, de même loi, selon une variable générique X admettant un vecteur d'espérance finie $\mathbb{E}\,[X]$ de dimension d et une matrice de variance-covariance* $\mathrm{Var}\,(X)$*. On a alors la convergence en loi dans $\mathbb{R}^d$*

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\,[X]\right) \to Z, \text{ où } Z \sim \mathcal{N}_d(0, \mathrm{Var}\,(X))$$

Proposition 8.17 (Propriété MME). *Si $\theta = \mathbb{E}\,[X]$ alors l'estimateur des moments de θ est $\widetilde{\theta}_n = \overline{X}_n$. La moyenne empirique est un estimateur sans biais et convergent en moyenne quadratique de $\mathbb{E}\,[X]$.*

Proposition 8.18 (Variance empirique). *La variance empirique* $S_n^2 = 1/n \sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2$ *est un estimateur asymptotiquement sans biais de* $\mathrm{Var}\,(X)$. *La variance estimée empirique* $S_n'^2 = 1/(n-1) \sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2$ *est un estimateur sans biais et convergent en moyenne quadratique de* $\mathrm{Var}\,(X)$.

Proposition 8.19 (Propriété MLE). *Sous (H1) à (H4), si les variables* $X_1, \ldots, X_n$ *sont indépendantes, de même loi dépendant d'un paramètre réel* θ, *alors l'estimateur MLE* $\widehat{\theta}_n$ *vérifie*

- $\left(\widehat{\theta}_n\right)_{n\geqslant 1}$ *converge presque sûrement vers* θ.
- $\left(\sqrt{I_n(\theta)}\left(\widehat{\theta}_n - \theta\right)\right)_{n\geqslant 1}$ *converge en loi vers la loi normale* $\mathcal{N}(0,1)$.

Proposition 8.20 (Méthode delta). *Sous (H1) à (H4), si* $\widehat{\theta}_n$ *est l'estimateur MLE de* θ *et* φ *est une fonction réelle dérivable alors* $\varphi(\widehat{\theta}_n)$ *est l'estimateur MLE de* $\varphi(\theta)$. *De plus, la suite* $\left(\sqrt{I_n(\theta)}\left(\varphi(\widehat{\theta}_n) - \varphi\left(\theta\right)\right)\right)_{n\geqslant 1}$ *converge en loi vers* $\mathcal{N}(0, \varphi'(\theta)^2)$ *si* $\varphi'(\theta) \neq 0$.

Proposition 8.21 (Propriété MLE). *Sous (H1) à (H4), si les variables* $X_1, \ldots, X_n$ *sont indépendantes de même loi dépendant d'un paramètre* $\theta \in \mathbb{R}^d$, *alors l'estimateur MLE* $\widehat{\theta}_n$ *vérifie*

- $\left(\widehat{\theta}_n\right)_{n\geqslant 1}$ *converge presque sûrement vers* θ.
- $\left(\sqrt{I_n(\theta)}\left(\widehat{\theta}_n - \theta\right)\right)_{n\geqslant 1}$ *converge en loi vers un vecteur aléatoire de taille* d *de loi normale centrée réduite.*

9 Vecteurs gaussiens

9.1 Définition et propriétés

Définition 9.1. *Un vecteur aléatoire* X *à valeurs dans* $\mathbb{R}^d$ *est un vecteur gaussien ssi toute combinaison linéaire de ses composantes est une variable réelle gaussienne.*

Proposition 9.2. *Le vecteur aléatoire* X *à valeurs dans* $\mathbb{R}^d$ *est un vecteur gaussien si et seulement si il existe un vecteur* $\mu \in \mathbb{R}^d$ *et une matrice* V *de taille* $d \times d$ *symétrique positive* ($V^t = V$ *et* $\langle x, Vx\rangle \geqslant 0, \forall x \in \mathbb{R}^d$) *tels que :*

$$\psi_X(u) = \exp\left(i\,\langle \mu, u\rangle - \frac{\langle u, Vu\rangle}{2}\right), \forall u \in \mathbb{R}^d.$$

De plus le vecteur X *est de carré intégrable et on a* $\mu = \mathbb{E}\,[X]$ *et* $V = \mathrm{Cov}(X, X)$. *Enfin, pour tout* $a \in \mathbb{R}^d$, *la loi de la variable* $\langle a, X\rangle$ *est la loi gaussienne* $\mathcal{N}(\langle a, \mu\rangle, \langle a, Va\rangle))$.

Proposition 9.3. *Soit* (X, Y) *un vecteur gaussien. L'équivalence suivante a lieu :*

$$\textit{les variables } X \textit{ et } Y \textit{ sont indépendantes} \iff \mathrm{Cov}(X, Y) = 0.$$

Proposition 9.4. *On considère une transformation affine de $\mathbb{R}^d$ dans $\mathbb{R}^n$: $x \longmapsto Mx + T$, où M est une matrice déterministe de taille $n \times d$ et T un vecteur déterministe de $\mathbb{R}^n$. Soit X un vecteur gaussien à valeurs dans $\mathbb{R}^d$ et de loi $\mathcal{N}(\mu, V)$. La variable aléatoire $MX + T$ est un vecteur gaussien à valeurs dans $\mathbb{R}^n$. De plus sa loi est*

$$\mathcal{L}(MX + T) = \mathcal{N}(T + M\mu, MVM').$$

Proposition 9.5. *Soit X un vecteur gaussien de $\mathbb{R}^d$ non dégénéré ($\det V > 0$) de loi $\mathcal{N}(\mu, V)$. Alors la matrice V est inversible et la loi de X possède la densité f : pour $x \in \mathbb{R}^d$:*

$$f_X(x) = \frac{1}{(2\pi)^{d/2}(\det V)^{1/2}} \exp\left(-\frac{\langle x - \mu, V^{\leftarrow}(x - \mu)\rangle}{2}\right).$$

9.2 Théorème central limite vectoriel

Proposition 9.6. *Soit $(X_n)_{n\geqslant 1}$ une suite de variables aléatoires à valeurs dans $\mathbb{R}^d$, indépendantes de même loi et de carré intégrable. On pose $\mu = \mathbb{E}[X_1]$, $V = \mathrm{Cov}(X_1, X_1) \in \mathbb{R}^{d\times d}$ et $\overline{X}_n = \frac{1}{n}\sum_{k=1}^n X_k$ la moyenne empirique. Alors la suite de vecteurs aléatoires $\left(\sqrt{n}(\overline{X}_n - \mu)\right)_{n\geqslant 1}$ converge en loi vers le vecteur gaussien de loi $\mathcal{N}(0, V)$:*

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, V).$$

Proposition 9.7. *Soit $(X_k)_{k\geqslant 1}$ une suite de variables à valeurs dans $\mathbb{R}^d$ indépendantes de même loi et de carré intégrable. On pose $\mu = \mathbb{E}[X_1]$, $V = \mathrm{Cov}(X_1, X_1) \in \mathbb{R}^{d\times d}$ et $\overline{X}_n = \frac{1}{n}\sum_{k=1}^n X_k$ la moyenne empirique. Soit g une fonction mesurable de $\mathbb{R}^d$ dans $\mathbb{R}^p$ continue et différentiable en μ. Sa différentielle au point μ est la matrice $\frac{\partial g}{\partial x}(\mu)$ de taille $p \times d$ définie par :*

$$\left(\frac{\partial g}{\partial x}(\mu)\right)_{i,j} = \frac{\partial g_i}{\partial x_j}(\mu), 1 \leqslant i \leqslant p, 1 \leqslant j \leqslant d.$$

On pose $\Sigma = \frac{\partial g}{\partial x}(\mu) V \left(\frac{\partial g}{\partial x}(\mu)\right)^t$. On a alors :

$$g(\overline{X}_n) \xrightarrow{p.s.} g(\mu) \quad et \quad \sqrt{n}\left(g(\overline{X}_n) - g(\mu)\right) \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

9.3 Interprétation géométrique de l'espérance conditionnelle

Proposition 9.8. *Supposons Y telle que $\mathbb{E}[Y^2] < \infty$. Parmi tous les réels a, la quantité $\mathbb{E}\left[(Y - a)^2\right]$ est minimale quand $a = \mathbb{E}[Y]$, i.e.,*

$$\min_{a\in\mathbb{R}} \mathbb{E}\left[(Y - a)^2\right] = \mathbb{E}\left[(Y - \mathbb{E}[Y])^2\right] = \mathrm{Var}(Y).$$

Définition 9.9. *La courbe $x \mapsto \mathbb{E}[Y|X = x]$ est appelée courbe de régression de Y en X.*

Théorème 9.10. *Supposons Y telle que $\mathbb{E}[Y^2] < \infty$. Parmi toutes les fonctions $u\colon \mathbb{R} \to \mathbb{R}$, l'erreur d'approximation $\mathbb{E}\left[(Y - u(X))^2\right]$ est minimale lorsque u est la fonction de régression $x \mapsto \mathbb{E}[Y|X = x]$, i.e., quand $u(X) = \mathbb{E}[Y|X]$.*

Définition 9.11. *La quantité*

$$\sigma^2 = \min_u \mathbb{E}\left[(Y - u(X))^2\right] = \mathbb{E}\left[(Y - \mathbb{E}[Y|X])^2\right]$$

est appelée erreur quadratique moyenne ou variance résiduelle.

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé. On note $L^2(\Omega) = L^2(\Omega, \mathcal{F}, \mathbb{P})$ l'ensemble des variables $X : \Omega \to \mathbb{R}$ de carré intégrable, i.e., $\mathbb{E}[X^2] < \infty$. On définit le produit scalaire dans $L^2(\Omega)$ par l'application

$$\langle \cdot, \cdot \rangle : L^2(\Omega) \times L^2(\Omega) \to \mathbb{R}$$
$$(X, Y) \mapsto \langle X, Y \rangle = \mathbb{E}[XY]$$

est un produit scalaire sur $L^2(\Omega)$ avec comme norme associée : $\|X\| = \sqrt{\mathbb{E}[X^2]}$.

Théorème 9.12. *(Théorème de projection orthogonale :) Soit H un sous espace fermé de $L^2(\Omega)$. Pour tout Y de $L^2(\Omega)$, il existe une unique variable de H, notée $\Pi_H(Y)$ qui soit à plus courte distance de Y. On l'appelle le projeté orthogonal de Y sur H et elle est entièrement caractérisée par la double propriété*

$$\Pi_H(Y) \in H$$
$$Y - \Pi_H(Y) \perp H.$$

10 Conditionnement des vecteurs gaussiens

Théorème 10.1. *Soit $(X, Y)^\top$ un vecteur gaussien, alors :*

$$\mathbb{E}[Y|X] = \mathbb{E}[Y] + \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}(X - \mathbb{E}[X]) = \widehat{a}X + \widehat{b}.$$

Autrement dit, la courbe de régression et la droite de régression coïncident.

On ne fait ici aucune hypothèse de gaussianité. On suppose observer n variables aléatoires $X_1, \ldots, X_n$ et on veut connaître la fonction affine des X_i donc de la forme

$$f(X_1, \ldots, X_n) = b + a_1 X_1 + \cdots + a_n X_n$$

qui approche le mieux la variable aléatoire Y au sens des moindres carrés, i.e., l'erreur quadratique moyenne

$$\mathbb{E}\left[(Y - (b + a_1 X_1 + \cdots + a_n X_n))^2\right]$$

soit minimale. Autrement dit, au lieu de chercher la droite de régression, on cherche l'hyperplan de régression. Ceci revient à déterminer la projection $\Pi_H(Y)$ de Y sur le sous-espace

$$H = \mathrm{Vect}(1, X_1, \ldots, X_n)$$

engendré par la constante 1 et les variables X_i.

Hypothèses :

- Notons $X = (X_1, \ldots, X_n)^\top$ le vecteur formé des variables X_i. On suppose que la matrice de dispersion $\Gamma_X = \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top\right]$ est inversible.
- Puisqu'on parle de projections et d'erreurs quadratiques, on suppose que toutes les variables aléatoires sont de carré intégrable.

Théorème 10.2 (Hyperplan de régression)**.** *La projection orthogonale de Y sur H est*

$$\Pi_H(Y) = \mathbb{E}[Y] + \Gamma_{Y,X}\Gamma_X^{-1}(X - \mathbb{E}[X])$$

avec

$$\Gamma_{Y,X} = \mathbb{E}\left[(Y - \mathbb{E}[Y])(X - \mathbb{E}[X])^\top\right] = [\mathrm{Cov}(Y, X_1), \ldots, \mathrm{Cov}(Y, X_n)],$$

matrice ligne de covariance de la variable aléatoire Y et du vecteur aléatoire X.

Corollaire 10.3. *L'erreur quadratique moyenne, encore appelée variance résiduelle ou résidu est*

$$\mathbb{E}\left[(Y - \Pi_H(Y))^2\right] = \Gamma_Y - \Gamma_{Y,X}\Gamma_X^{-1}\Gamma_{X,Y}$$

avec $\Gamma_Y = \mathrm{Var}(Y)$ et $\Gamma_{X,Y} = (\Gamma_{Y,X})'$.

On a vu que pour un vecteur gaussien bidimensionnel $(X \quad Y)^\top$ la droite de régression coincide avec la courbe de régression. Plus généralement, on montre que pour un vecteur gaussien $(X_1, \ldots, X_n, Y)^\top$, l'espérance conditionnelle coïncide avec la projection sur l'hyperplan de régression.

Théorème 10.4 (Espérance conditionnelle $\Leftrightarrow$ hyperplan de régression)**.** *Si $(X_1, \ldots, X_n, Y)^\top$ est gaussien alors*

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|X_1, \ldots, X_n] = \mathbb{E}[Y] + \Gamma_{Y,X}\Gamma_X^{-1}(X - \mathbb{E}[X])$$

et la variance résiduelle

$$\sigma^2 = \mathbb{E}\left[(Y - \mathbb{E}[Y|X])^2\right] = \Gamma_Y - \Gamma_{Y,X}\Gamma_X^{-1}\Gamma_{X,Y}.$$

On a obtenu la décomposition indépendante :

$$Y = \mathbb{E}[Y|X] + W = \big(\mathbb{E}[Y] + \Gamma_{Y,X}\Gamma_X^{-1}(X - \mathbb{E}[X])\big) + W$$

avec $W = Y - \mathbb{E}[Y|X]$ qui est

- une variable aléatoire gaussienne car combinaison linéaire de X et Y qui forme un vecteur gaussien. En effet $W = Y - \mathbb{E}[Y] - \Gamma_{Y,X}\Gamma_X^{-1}(X - \mathbb{E}[X])$;
- centrée car $\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$;
- indépendante des X_i car $\mathrm{Cov}(W, X_i) = \mathbb{E}[WX_i] = \mathbb{E}[YX_i] - \mathbb{E}[\mathbb{E}[YX_i \mid X]] = 0$ et le vecteur formé par W et X_i est gaussien ;
- sa variance résiduelle est $\sigma^2 = \Gamma_Y - \Gamma_{Y,X}\Gamma_X^{-1}\Gamma_{X,Y}$.

Donc, on a

$$W \sim \mathcal{N}(0, \sigma^2)$$
$$W \perp X.$$

Ainsi la loi de Y sachant $\{X = x\}$ est

$$Y|_{\{X=x\}} = \mathbb{E}[Y|X = x] + W \sim \mathcal{N}\left(\mathbb{E}[Y|X = x], \sigma^2\right).$$

11 Régression linéaire

Définition 11.1. *Le modèle de régression de Y sur x est défini par*

$$Y = f(x) + \varepsilon$$

où

- *Y est la variable à expliquer ou variable expliquée*
- *x est la variable explicative ou prédicteur ou régresseur*
- *ε est l'erreur de prévision de Y par $f(x)$ ou résidu.*

Définition 11.2. *Le modèle de régression linéaire simple ou modèle linéaire simple est défini par*

$$Y_i = \beta_1 x_i + \beta_0 + \varepsilon_i \quad i = 1, \ldots, n$$

où β_0 et β_1 sont des paramètres réels inconnus, et les ε_i sont indépendants, de même loi, centrés et de variance σ^2.

On cherchera une prévision $\widehat{Y}$ tel que :

$\mathbb{E}\left[Y - \widehat{Y}\right] = 0$ prévision "sans biais"

$\mathrm{Var}\left(Y - \widehat{Y}\right)$ soit la plus petite possible : faible variance de l'erreur de prévision.

Le problème de la prévision de Y à l'aide de X est parfaitement résolu si on se place dans l'espace L^2 des variables réelles de carré intégrable :

$$L^2 = \{\text{variables aléatoires réelles } X : \mathbb{E}\left[X^2\right] < \infty\}.$$

En effet, dans cet espace, il est facile de voir que la forme bilinéaire $\langle X, Y\rangle = \mathbb{E}[XY]$ est un produit scalaire, et que L^2 muni de ce produit scalaire est un espace de Hilbert. La norme associée est $\|X\| = \sqrt{\mathbb{E}[X^2]}$.

Proposition 11.3. *On a*

- $\|X - \mathbb{E}[X]\|^2 = \mathrm{Var}(X)$
- $\langle X - \mathbb{E}[X], Y - \mathbb{E}[Y]\rangle = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathrm{Cov}(X, Y)$.

Théorème 11.4. *La projection orthogonale de X sur l'ensemble des variables aléatoires constantes est $\mathbb{E}[X]$. En d'autres termes, la meilleure approximation (au sens de la norme dans L^2) d'une variable X par une constante est son espérance $\mathbb{E}[X]$.*

Considérons maintenant l'ensemble L^2_X des variables fonction de X

$$L^2_X = \{f(X); f : \mathbb{R} \longrightarrow \mathbb{R}\}.$$

On montre que L^2_X est un sous-espace vectoriel fermé de L^2. Trouver la meilleure prévision $\widehat{Y} = f(X)$ de Y, c'est trouver l'élément de L^2_X le plus proche de Y au sens où $\text{Var}\left(Y - \widehat{Y}\right)$ est minimum, sous la contrainte que $\mathbb{E}\left[Y - \widehat{Y}\right] = 0$. Or

$$\text{Var}\left(Y - \widehat{Y}\right) = \mathbb{E}\left[Y - \widehat{Y}\right]^2 - \left(\mathbb{E}\left[Y - \widehat{Y}\right]\right)^2 = \|Y - \widehat{Y}\|^2.$$

On sait que le minimum des $\|Y - f(Y)\|^2$ est réalisé quand $f(X)$ est la projection orthogonale de Y sur L^2_X. Par conséquent, la prévision optimale cherchée est cette projection orthogonale, qui n'est autre que l'espérance conditionnelle de Y sachant X. D'où :

Théorème 11.5. *La projection orthogonale de Y sur L^2_X est $\mathbb{E}[Y|X]$. En d'autres termes, la meilleure approximation (au sens de la norme dans L^2) de Y par une fonction de X est $\mathbb{E}[Y|X]$.*

Théorème 11.6. *Les estimateurs des moindres carrés de β_1 et β_0 sont*

$$\widehat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x}_n)(Y_i - \overline{Y}_n)}{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x}_n)^2} \quad et \quad \widehat{\beta}_0 = \overline{Y}_n - \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x}_n)(Y_i - \overline{Y}_n)}{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x}_n)^2}\,\overline{x}_n.$$

Théorème 11.7. *$\widehat{\beta}_1$ et $\widehat{\beta}_0$ sont des ESB de β_1 et β_0.*

Remarque 11.8. Ces estimateurs $\widehat{\beta}_1$ et $\widehat{\beta}_0$ peuvent simplement se réécrire sous la forme

$$\widehat{\beta}_1 = \frac{C_{xY}}{s_x^2} \quad \text{et} \quad \widehat{\beta}_0 = \overline{Y}_n - \widehat{\beta}_1 \overline{x}_n$$

où C_{xY} désigne la covariance empirique entre les x_i et les Y_i.

Théorème 11.9. (Gauss Markov). *$\widehat{\beta}_1$ et $\widehat{\beta}_0$ sont les estimateurs de β_1 et β_0 sans biais et de variance minimale (ESBVM) parmi tous les ESB (combinaisons linéaires des Y_i.)*

Il reste maintenant à estimer la variance du bruit σ^2. Puisque, $\forall i$, $\sigma^2 = \text{Var}(\varepsilon_i) = \text{Var}(Y_i - \beta_1 x_i - \beta_0)$, il est naturel d'estimer σ^2 par la variance empirique des $Y_i - \widehat{\beta}_1 x_i - \widehat{\beta}_0$.

Définition 11.10. *Les variables*

- *$E_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\beta}_1 x_i - \widehat{\beta}_0$, $i = 1, \ldots, n$ sont appelées les résidus empiriques*
- *La variance empirique des résidus empiriques est notée $S^2_{Y|x} = \frac{1}{n}\sum_{i=1}^n E_i^2 - \overline{E}_n^2$ et est appelée variance résiduelle.*

Proposition 11.11. *$\widehat{\sigma}^2 = \frac{n}{n-2} S^2_{Y|x}$ est un ESB de σ^2.*

On remarque que, par définition des E_i, on a $\overline{E}_n = \overline{Y}_n - \widehat{\beta}_1 \overline{x}_n - \widehat{\beta}_0$ qui vaut 0 par définition de $\widehat{\beta}_0$. Donc $S^2_{Y|x} = \frac{1}{n}\sum_{i=1}^n E_i^2$ et par conséquant $\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n E_i^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - \widehat{\beta}_1 x_i - \widehat{\beta}_0)^2$ est un estimateur sans biais de σ^2.

Définition 11.12. *Le modèle de régression linéaire simple gaussien est défini par*

$$Y_i = \beta_1 x_i + \beta_0 + \varepsilon_i \quad i = 1, \ldots, n$$

où les ε_i sont des variables aléatoires indépendantes et de même loi normale centrée et de variance σ^2, $\mathcal{N}(0, \sigma^2)$.

Proposition 11.13. *En notant s_x^2 la variance empirique des x_i, on a*

- $\forall i \in \{1, \ldots, n\}$, $Y_i \sim \mathcal{N}(\beta_1 x_i + \beta_0, \sigma^2)$
- $\widehat{\beta}_1 \sim \mathcal{N}\Big(\beta_1, \sigma^2/(n s_x^2)\Big)$
- $\widehat{\beta}_0 \sim \mathcal{N}\Big(\beta_0, \frac{\sigma^2}{n}(1 + \frac{\overline{x}_n^2}{s_x^2})\Big)$
- $\frac{(n-2)\widehat{\sigma}^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \widehat{\beta}_1 x_i - \widehat{\beta}_0)^2 \sim \chi^2_{n-2}$
- *$\widehat{\sigma}^2$ est indépendant de $\overline{Y}_n$, $\widehat{\beta}_1$ et $\widehat{\beta}_0$.*

Théorème 11.14. *Dans le modèle linéaire simple gaussien, les estimateurs du maximum de vraisemblance de β_1, β_0 et σ^2 sont $\widehat{\beta}_1$, $\widehat{\beta}_0$ et $\frac{n-2}{n}\widehat{\sigma}^2$.*

Définition 11.15. *On appelle loi de Student à n degrés de liberté, notée $T(n)$, la variable aléatoire de la forme*

$$\frac{Z}{\sqrt{W/n}},$$

où Z est de loi $\mathcal{N}(0,1)$, W est de loi χ^2_n et Z et W sont indépendantes, noté $Z \perp W$.
On appelle loi de Fisher à n et m degrés de liberté, notée $F(n,m)$, la variable aléatoire de la forme

$$\frac{W/n}{Y/m},$$

où W est de loi χ^2_n, Y est de loi χ^2_m et $W \perp Y$.

Proposition 11.16.
1. *Si $X \sim \mathcal{N}(0,1)$, alors $X^2 \sim \chi^2_1 = G(\frac{1}{2}, \frac{1}{2})$.*
2. *Si $Y_1, \ldots, Y_n$ sont indépendantes et de même loi χ^2_1, alors $\sum_{i=1}^n Y_i \sim \chi^2_n = G(\frac{n}{2}, \frac{1}{2})$.*
3. *$X_1, \ldots, X_n$ indépendantes de loi $\mathcal{N}(m, \sigma^2)$, alors :*
   (a) *$\forall i$, $\frac{X_i - m}{\sigma} \sim \mathcal{N}(0,1)$,*
   (b) *$\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - m)^2 \sim \chi^2_n$,*

(c) $\overline{X}_n \sim \mathcal{N}(m, \sigma^2/n)$

(d) $\sqrt{n}\,\frac{\overline{X}_n - m}{\sigma} \sim \mathcal{N}(0,1)$,

(e) $\sqrt{n}\,\frac{\overline{X}_n - m}{s'_n} \sim T(n-1)$ *où* $s'^2_n = \frac{n}{n-1}s^2_n$.,

(f) $\frac{ns_n^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \overline{X}_n)^2 \sim \chi^2_{n-1}$,

(g) $\overline{X}_n$ *et* s_n^2 *sont indépendants (cf. exercice 3, TD8).*

En utilisant ces propriétés, on montre facilement que

$$\sqrt{n}\,s_x\,\frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}} \sim T(n-2) \quad \text{et} \quad \frac{\widehat{\beta}_0 - \beta_0}{\widehat{\sigma}}\,\frac{\sqrt{n}\,s_x}{\sqrt{s_x^2 + \overline{x}_n^2}} \sim T(n-2).$$

Proposition 11.17 (Intervalles de confiance)**.** *On notera* $t_{n-2,1-\alpha}$ *le quantile d'ordre* $1-\alpha$ *de la loi* $T(n-2)$*. Alors*

- *L'intervalle*
$$\left[\widehat{\beta}_1 - \frac{\widehat{\sigma}t_{n-2,1-\alpha/2}}{s_x\sqrt{n}}; \widehat{\beta}_1 + \frac{\widehat{\sigma}t_{n-2,1-\alpha/2}}{s_x\sqrt{n}}\right]$$
est un intervalle de confiance bilatéral de β_1*, au niveau de confiance* $1-\alpha$*.*

- *L'intervalle*
$$\left[\widehat{\beta}_0 - \frac{\widehat{\sigma}t_{n-2,1-\alpha/2}\sqrt{s_x^2 + \overline{x}_n^2}}{s_x\sqrt{n}}; \widehat{\beta}_0 + \frac{\widehat{\sigma}t_{n-2,1-\alpha/2}\sqrt{s_x^2 + \overline{x}_n^2}}{s_x\sqrt{n}}\right]$$
est un intervalle de confiance bilatéral de β_0*, au niveau de confiance* $1-\alpha$*.*

- *Un IC de seuil* α *pour* σ^2 *est*
$$\left[\frac{(n-2)\widehat{\sigma}^2}{z_{n-2,\frac{\alpha}{2}}}, \frac{(n-2)\widehat{\sigma}^2}{z_{n-2,1-\frac{\alpha}{2}}}\right],$$
où $z_{n,\alpha}$ *est le quantile de la* χ^2*.*

Proposition 11.18. *Tests d'hypothèse bilatéraux sur* β_1, β_0 *et* σ^2*, en notant* W *la région critique,*

- *Test de seuil* α *de "*$\beta_1 = b$*" contre "*$\beta_1 \neq b$*"*
$$W = \left\{(y_1, \ldots, y_n); \left|\frac{\widehat{\beta}_1 - b}{\widehat{\sigma}} s_x\sqrt{n}\right| > t_{n-2,\alpha}\right\}.$$

- *Test de seuil* α *de "*$\beta_0 = b$*" contre "*$\beta_0 \neq b$*"*
$$W = \left\{(y_1, \ldots, y_n); \left|\frac{\widehat{\beta}_0 - b}{\widehat{\sigma}}\,\frac{s_x\sqrt{n}}{\sqrt{s_x^2 + \overline{x}_n^2}}\right| > t_{n-2,\alpha}\right\}.$$

- *Test de seuil* α *de "*$\sigma = \sigma_0$*" contre "*$\sigma \neq \sigma_0$*"*
$$W = \left\{\frac{(n-2)\widehat{\sigma}^2}{\sigma_0^2} < z_{n-2,1-\frac{\alpha}{2}} \quad ou \quad \frac{(n-2)\widehat{\sigma}^2}{\sigma_0^2} > z_{n-2,\frac{\alpha}{2}}\right\}.$$

3. Test de pertinence de la régression linéaire La région critique d'un tel test est

$$R = \left\{ \left| \frac{\widehat{\beta}_1}{\widehat{\sigma}} s_x \sqrt{n} \right| > t_{n-2,1-\alpha/2} \right\}.$$

On peut l'écrire de façon plus parlante en faisant intervenir le coefficient de corrélation empirique. En effet, on a

$$\widehat{\beta}_1 = r_{xY} \frac{s_Y}{s_x} \quad \text{et} \quad \widehat{\sigma}^2 = \frac{n}{n-2} s_Y^2 \left(1 - r_{xY}^2\right).$$

Par conséquent

$$\frac{\widehat{\beta}_1}{\widehat{\sigma}} s_x \sqrt{n} = \frac{r_{xY}}{\sqrt{1 - r_{xY}^2}} \sqrt{n-2},$$

et donc R se réécrit comme

$$R = \left\{ \sqrt{n-2} \frac{|r_{xY}|}{\sqrt{1 - r_{xY}^2}} > t_{n-2,1-\alpha/2} \right\}.$$

Remarque 11.19. La région critique du test

$$H_0 : \beta_1 = 0 \quad \text{contre} \quad H_1 : \beta_1 \neq 0$$

est

$$R = \left\{ \left| \frac{\widehat{\beta}_1}{\widehat{\sigma}} s_x \sqrt{n} \right| > t_{n-2,\alpha} \right\}$$

donc si R est vraie, cela veut dire qu'on rejette H_0, i.e., $\beta_1 \neq 0$, i.e., $r_{xY} \neq 0$, donc on accepte la linéarité.