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

Davide Giraudo

25 mars 2026

Table des matières

1 Outils probabilistes

Dans cette section, $(\Omega, \mathcal{F}, \mathbb{P})$ désigne un espace probabilisé.

1.1 Lois et densités usuelles

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**.

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	$n \in \mathbb{N}^*, p \in [0, 1]$	$\{0, \ldots, n\}$	$\mathbb{P}(X = i) = C_n^i \, p^i \, (1 - p)^{n-i}$
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 1.1. *Soit X une variable réelle. On dit que X a pour densité f_X si pour tout borélien B,*

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

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 – Densités usuelles

- **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 $\mathbb{E}\left[\varphi(X)\right]$ existe alors on a l'inégalité suivante :

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

- **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

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

Définition 1.2. *La loi Gamma de paramètres (λ, α) admet comme densité*

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \mathbf{1}_{\mathbb{R}_+}(x),$$

où $\Gamma(\cdot)$ est définie par

$$\Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1} e^{-t} dt.$$

1.2 Vecteurs aléatoires

Définition 1.3. *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 1.4. *On dit que $X = (X_1, \ldots, X_N)$ a pour densité f_X si pour tous $x_1, \ldots, x_N \in \mathbb{R}$,*

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

Proposition 1.5. *Si $X = (X_1, \ldots, X_N)$ a pour densité f_X, alors λ-presque partout,*

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

Théorème 1.6. *Soit $X = (X_1, \ldots, X_N)$ ayant pour densité f_X. Si une fonction $g : \mathbb{R}^N \to \mathbb{R}$ est telle que*

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

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

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

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

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

Définition 1.7. *Soit $X = (X_1, \ldots, X_N)$ un vecteur aléatoire. Si $X_1, \ldots, X_N$ admettent toutes un moment d'ordre 1, alors le vecteur*

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

est appelé l'espérance de X.

Définition 1.8. *Soient X et Y deux variables réelles admettant toutes des moments d'ordre 2. Alors*

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

est appelée la covariance entre X et Y.

Proposition 1.9. *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 1.10. *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 1.11. *Soit $X = (X_1, \dots X_N)$ un vecteur aléatoire telle que chaque composante admet un moment d'ordre 2. On appelle*

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

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 1.12. *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 1.13. *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 $X_1, \dots, X_N, Y_1, \dots, Y_M$ sont des variables réelles admettant toutes un moment d'ordre 2, alors pour tous réels $a_1, \dots, a_N, b_1, \dots, b_M$:*

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

Théorème 1.14. *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

$$f_Y(y) = f_X(h^{\leftarrow}(y))\, |\det(Dh^{\leftarrow})(h^{\leftarrow}(y))|$$

où $(Dh^{\leftarrow})(x)$ est le jacobien de $h^{\leftarrow}$ en x : soit $h^{\leftarrow} = (h_1^{\leftarrow}, \dots, h_N^{\leftarrow})$

$$(Dh^{\leftarrow})(x) = \begin{pmatrix} \frac{\partial h_1^{\leftarrow}(x_1,\dots,x_N)}{\partial x_1} & \dots & \frac{\partial h_1^{\leftarrow}(x_1,\dots,x_N)}{\partial x_N} \\ \vdots & \dots & \vdots \\ \vdots & \dots & \vdots \\ \vdots & \dots & \vdots \\ \frac{\partial h_N^{\leftarrow}(x_1,\dots,x_N)}{\partial x_1} & \dots & \frac{\partial h_N^{\leftarrow}(x_1,\dots,x_N)}{\partial x_N} \end{pmatrix}$$

Définition 1.15. *Soit $X = (X_1, \dots, X_N)$ un vecteur aléatoire. On appelle loi marginale la loi de tout sous-vecteur $(X_{i_1}, \dots, X_{i_k})$ (avec $k < N$) extrait de X.*

Proposition 1.16. *Soit*

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

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

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

On écrit souvent

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

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

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

Proposition 1.17. *Si $f_{(X_1,\dots,X_n)}$ est la densité du vecteur aléatoire $X = (X_1, \dots, X_n)$, alors le vecteur aléatoire $(X_1, \dots, X_k)$ admet la densité*

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

Définition 1.18. *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 1.19. *Soient X et Y deux variables aléatoires. On suppose que X est intégrable. L'espérance conditionnelle de X sachant Y, notée, $\mathbb{E}[X \mid Y]$, est définie comme étant l'unique variable aléatoire Z s'exprimant comme une fonction de Y et vérifiant pour tout borélien B de $\mathbb{R}$ l'égalité $\mathbb{E}[X \mathbf{1}_{Y \in B}] = \mathbb{E}[Z \mathbf{1}_{Y \in B}]$.*

Interprétation lorsque X a un moment d'ordre 2 : c'est la variable aléatoire la plus proche de X qui s'exprime comme une fonction de Y.

Notons que $\mathbb{E}[X \mid Y] = h(Y)$ avec h donnée par

$$h(y) = \int_{\mathbb{R}} f_{X|\{Y=y\}}(x) dx. \tag{1.1}$$

Définition 1.20. *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 $A \in \mathcal{B}(\mathbb{R}^d)$ et tout $B \in \mathcal{B}(\mathbb{R}^{d'})$,*

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

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

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

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

Proposition 1.21. *Pour que deux variables aléatoires X et Y soient indépendantes, il faut et il suffit que*

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

pour toute fonction bornée $f : \mathbb{R}^d \to \mathbb{R}$ et toute fonction bornée $g : \mathbb{R}^{d'} \to \mathbb{R}$.

Définition 1.22. *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

$$\mathbb{E}\left[f_1(X_1) \dots f_n(X_n)\right] = \mathbb{E}\left[f_1(X_1)\right] \dots \mathbb{E}\left[f_n(X_n)\right]$$

pour toutes les fonctions bornées $f_k : E_k \to \mathbb{R}$.

Théorème 1.23. *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 1.24. *Soit $(X_1, \dots, X_n)$ un vecteur aléatoire à densité. Alors $X_1, \dots, X_n$ sont indépendantes ssi*

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

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

Proposition 1.25. *Si la densité du vecteur aléatoire $(X_1, \dots, X_n)$ se factorise*

$$f_{(X_1,\dots,X_n)}(x_1, \dots, x_n) = g_1(x_1) \dots g_n(x_n)$$

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

Définition 1.26. *Soit X une variable aléatoire à valeurs dans $\mathbb{R}^N$ muni du produit scalaire $\langle t, X \rangle = \sum_{j=1}^N t_j X_j$. L'application $\varphi_X : \mathbb{R}^N \to \mathbb{C}$ donnée par*

$$\varphi_X(t) = \mathbb{E}\left[e^{i\langle t, X\rangle}\right]$$

s'appelle la fonction caractéristique de X.

Théorème 1.27. *Deux variables aléatoires X et Y ont même loi ssi $\varphi_X = \varphi_Y$. Autrement dit, comme son nom l'indique, la fonction caractéristique φ_X caractérise la loi de X.*

Théorème 1.28. *Les variables réelles $X_1, \dots, X_n$ sont indépendantes ssi*

$$\varphi_X(t) = \prod_{k=1}^n \varphi_{X_k}(t_k) \qquad \forall t = (t_1, \dots, t_n) \in \mathbb{R}^n$$

où $X := (X_1, \dots, X_n)$.

Proposition 1.29. *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*

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

En particulier

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

1.3 Théorèmes fondamentaux en probabilité

Définition 1.30. *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 1.31. *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 1.32. *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 1.33. *Si $X_n \to X$ en moyenne d'ordre p, alors la convergence a encore lieu en probabilité.*

Définition 1.34. *La suite $(X_n)_{n\geqslant 1}$ converge presque sûrement vers X si*

$$\mathbb{P}\left(\left\{\omega \in \Omega, \lim_n X_n(\omega) = X(\omega)\right\}\right) = 1.$$

Proposition 1.35. *Une suite $(X_n)_{n\geqslant 1}$ converge presque sûrement vers X si et seulement si la suite $\left(\sup_{k\geqslant n} |X_k - X|\right)_{n\geqslant 1}$ converge vers 0 en probabilité.*

Proposition 1.36. *Si pour tout $\varepsilon > 0$*

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

alors X_n converge presque sûrement vers X.

1.4 Convergence en loi

Définition 1.37. *On dit qu'une suite de variables $X_1, \ldots, X_n, \ldots$ 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 $\mathbb{R}^N$*

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

Proposition 1.38. *Si $X_1, \ldots, X_n, \ldots$ est une suite de variables réelles qui converge en probabilité vers une variable X alors X_n converge également vers X en loi.*

Théorème 1.39. *Soient $X_1, \ldots, X_n, \ldots$ 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 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 1.40. *Soit une suite de variables $X, X_1, \ldots, X_n, \ldots$ à valeurs dans $\mathbb{R}^N$. Les conditions suivantes sont équivalentes :*

1. *$(X_n)_{n\geqslant 1}$ converge en loi vers X.*
2. *$(\varphi_{X_n})_{n\geqslant 1}$ converge simplement vers φ_X.*

Liens entre les différents modes de convergence :
Convergences p.s. et L^p impliquent convergence proba
Convergence proba implique convergence en loi
Convergence en loi implique convergence proba quand la limite est constante
Convergence proba ou L^p impliquent convergence p.s. d'une sous-suite.

Théorème 1.41 (Théorème limite central)**.** *Soit $X_1, \ldots, X_n, \ldots$ 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 $\mathcal{N}(0, \sigma^2)$

Corollaire 1.42. *Soit $(X_n, n \in \mathbb{N}^*)$ une suite de variables réelles indépendantes, identiquement distribuées de carré intégrable. On pose $\sigma^2 = \mathrm{Var}(X_n)$. Alors pour $a > 0$:*

$$\mathbb{P}\left(\mathbb{E}[X] \in \left[\overline{X}_n - \frac{a\sigma}{\sqrt{n}}, \overline{X}_n + \frac{a\sigma}{\sqrt{n}}\right]\right) \overset{n\to\infty}{\longrightarrow} \int_{-a}^{a} e^{-x^2/2}\frac{dx}{\sqrt{2\pi}}.$$

2 Vecteurs gaussiens

2.1 Définition et propriétés

Définition 2.1. *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 (possiblement dégénérée).*

Proposition 2.2. *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 :*

$$\forall u \in \mathbb{R}^d, \quad \varphi_X(u) = \mathbb{E}\left[e^{i\langle u, X\rangle}\right] = \exp\left(i\langle \mu, u\rangle - \frac{\langle u, Vu\rangle}{2}\right).$$

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 2.3. *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*

$$\mathcal{L}(MX + T) = \mathcal{N}(T + M\mu, MVM^\top).$$

Proposition 2.4. *Soit X un vecteur gaussien de $\mathbb{R}^d$ non dégénéré ($\det V > 0$) de loi $\mathcal{N}(\mu, V)$. Alors la loi de X a pour densité f : pour $x \in \mathbb{R}^d$:*

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

Proposition 2.5. *Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes de loi normale centrée réduite. Alors*

$$\sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2$$

suit une loi du χ^2 à $n-1$ degrés de liberté, autrement dit, une loi gamma avec $\lambda = 1/2$ et $\alpha = n/2$.

Définition 2.6. *On appelle loi de Student à n degrés de liberté, notée $T(n)$, la loi d'une 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 2.7.

1. *Si $X \hookrightarrow \mathcal{N}(0,1)$, alors $X^2 \hookrightarrow \chi^2_1 = G(\frac{1}{2}, \frac{1}{2})$.*
2. *Si $Y_1, \ldots, Y_n$ sont indépendantes et de même loi χ^2_1, alors $\sum_{i=1}^n Y_i \hookrightarrow \chi^2_n = G(\frac{n}{2}, \frac{1}{2})$.*
3. *$X_1, \ldots, X_n$ indépendantes de loi $\mathcal{N}(m, \sigma^2)$, alors :*
   - *(a) $\forall i, \frac{X_i - m}{\sigma} \hookrightarrow \mathcal{N}(0,1)$,*
   - *(b) $\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - m)^2 \hookrightarrow \chi^2_n$,*
   - *(c) $\overline{X}_n \hookrightarrow \mathcal{N}(m, \sigma^2/n)$*
   - *(d) $\sqrt{n}\,\frac{\overline{X}_n - m}{\sigma} \hookrightarrow \mathcal{N}(0,1)$,*
   - *(e) $\sqrt{n}\,\frac{\overline{X}_n - m}{s'_n} \hookrightarrow T(n-1)$ où $s'^2_n = \frac{n}{n-1}s^2_n$.,*
   - *(f) $\frac{ns^2_n}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \overline{X}_n)^2 \hookrightarrow \chi^2_{n-1}$,*
   - *(g) $\overline{X}_n$ et s^2_n sont indépendants.*

Théorème 2.8 (Théorème limite central vectoriel)**.** *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.*

Soit g une fonction mesurable de $\mathbb{R}^d$ and $\mathbb{R}^p$ 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 :

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

En particulier, avec $p = d$ et $g(x) = x$,

$$\overline{X}_n \xrightarrow{p.s.} \mu$$

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 2.9. *Supposons Y telle que $\mathbb{E}[Y^2] < \infty$. Alors*

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

Définition 2.10. *La courbe $x \mapsto \mathbb{E}[Y|X = x]$ est appelée courbe de régression de Y en X.*

Théorème 2.11. *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 2.12. *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 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

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

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

Théorème 2.13 (Théorème de projection orthogonale)**.** *Soit H un sous espace fermé de $L^2(\Omega)$. $\forall Y$ de $L^2(\Omega)$, il existe une unique variable de H, notée $\Pi_H(Y)$ qui soit à plus courte distance de Y et appelée projeté orthogonal de Y sur H. Elle est entièrement caractérisée par la double propriété*

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

2.2 Conditionnement des vecteurs gaussiens

Théorème 2.14. *Soit $(X, Y)^\perp$ 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 suppose observer n variables aléatoires $X_1, \ldots, X_n$ et on veut connaître la fonction affine des X_i donc de la forme

$$f(X_1, \ldots, X_n) = b + a_1X_1 + \cdots + a_nX_n$$

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

$$\mathbb{E}\left[(Y - (b + a_1X_1 + \cdots + a_nX_n))^2\right]$$

soit minimale. 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 2.15 (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 2.16. *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 2.17 (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 :

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

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

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

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

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

Théorème 2.18 (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

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

Théorème 2.19 (Théorème limite central)**.** *Soit $(X_i)_i$ 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 $\mathbb{R}^d$*

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

On peut appliquer le théorème limite central vectoriel au test du χ-deux. On cherche à déterminer si les observations $x_1, \ldots, x_n$ proviennent de réalisations de variables aléatoires $X_1, \ldots, X_n$ qui suivent une loi discrète donnée par $\mathbb{P}(X = a_k) = p_k$, $1 \leqslant k \leqslant K$ et $a_1, \ldots, a_K$ sont les valeurs que peut prendre X_1. Par exemple, pour tester si une pièce est équilibrée, on prend $K = 2$, $p_1 = p_2 = 1/2$ et $a_1 =$ pile, $a_2 =$ face.

On compare les fréquences observées avec les fréquences attendues. On pose

$$N_k = \sum_{i=1}^{n} \mathbf{1}_{x_i = a_k}, \quad 1 \leqslant k \leqslant K$$

et

$$T(x_1, \ldots, x_n) = \sum_{k=1}^{K} \frac{(N_k - np_k)^2}{np_k}.$$

Proposition 2.20 (Fondement du test du χ-deux d'adéquation)**.** *Si pour tout $k \in \{1, \ldots, K\}$, $\mathbb{P}(X_1 = a_k) = p_k$, alors $T(X_1, \ldots, X_n)$ converge en loi vers une variable aléatoire de loi du χ-deux à $K-1$ degrés de liberté : pour tout $t \in \mathbb{R}$,*

$$\mathbb{P}(T(X_1, \ldots, X_n) \leqslant t) \to \mathbb{P}\left(\chi^2_{K-1} \leqslant t\right),$$

où χ^2_{K-1} est une variable aléatoire de loi du χ-deux à $K-1$ degrés de liberté.

S'il existe k tel que $\mathbb{P}(X_1 = a_k) \neq p_k$, alors $T(X_1, \ldots, X_n) \to +\infty$ presque sûrement.

3 Inférence statistique

Idée globale : on dispose d'observations $x_1, \ldots, x_n$ issues de la réalisation d'un vecteur aléatoire $(X_1, \ldots, X_n)$ qui a une loi dépendant d'un paramètre θ. On cherche à estimer ce paramètre à l'aide d'une quantité calculable à partir des observations et à avoir des informations sur la différence entre l'estimateur et le vrai paramètre que l'on ne connaît pas.

Définition 3.1. *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 3.2. *Soit* $\left(\mathbb{X},(P_\theta)_{\theta\in\Theta}\right)$ *un modèle statistique. On appelle statistique sur ce modèle toute application* $\varphi\colon \mathbb{X}\to\mathbb{Y}$ *indépendante de* θ*, mesurable par rapport aux tribus* $\mathcal{A}$ *et* $\mathcal{B}$ *sur* $\mathbb{X}$ *(respectivement* $\mathbb{Y}$*). Le modèle* $\left(\mathbb{Y},(Q_\theta)_{\theta\in\theta}\right)$ *où pour tout* θ*,* Q_θ *est la mesure image par* φ *de* P_θ*, est dit modèle image par* φ *du modèle* $\left(\mathbb{X},(P_\theta)_{\theta\in\Theta}\right)$.

Définition 3.3. *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 3.4. *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* $\theta\in\Theta$*,* $\mathbb{P}_\theta(A)=0$*. On dit que* μ *est une dominante du modèle et on appelle vraisemblance du modèle toute application* $L\colon \mathbb{X}\times\Theta\to\mathbb{R}_+$ *tel que pour tout* $\theta\in\Theta$*, l'application partielle* $L(\cdot,\theta):\mathbb{X}\to\mathbb{R}_+$ *soit une densité de* P_θ *par rapport à* μ*, c'est-à-dire*

$$\forall A\in\mathcal{A},\quad P_\theta(A)=\int L(x,\theta)\,d\mu(x) \tag{3.1}$$

La loi image d'une mesure μ par une application g est la mesure $B\mapsto\mu(g^{-1}(B))$.

Définition 3.5. *Un modèle réel unidimensionnnel* $\left(\mathbb{R},(P_\theta)_{\theta\in\mathbb{R}}\right)$ *est dit à paramètre de position si pour tout* $\theta\in\mathbb{R}$*, la loi* P_θ *est l'image de la loi* P_0 *par la translation* $x\mapsto x+\theta$*. On dit alors que le paramètre inconnu* θ *est un paramètre de position.*

Définition 3.6. *Un modèle réel multidimensionnnel* $\left(\mathbb{R}^n,(P_\theta)_{\theta\in\mathbb{R}^n}\right)$ *est dit à paramètre de position si pour tout* $\theta\in\mathbb{R}^n$*, la loi* P_θ *est l'image de la loi* P_0 *par la translation* $x\mapsto x+\theta$*. On dit alors que le paramètre inconnu* θ *est un paramètre de position.*

Définition 3.7. *Un modèle réel multidimensionnnel* $\left(\mathbb{R}^n,(P_\theta)_{\theta\in\mathbb{R}_+^*}\right)$ *est dit à paramètre d'échelle si pour tout* $\theta\in\mathbb{R}_+^*$*, la loi* P_θ *est l'image de la loi* P_1 *par l'homothétie* $x\mapsto\theta x$ *de rapport* θ*. On dit alors que le paramètre inconnu* θ *est un paramètre d'échelle.*

4 Exhaustivité

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. Souvent, ce n'est pas l'observation x elle-même qui est considérée mais une certaine fonction de cette donnée, soit $\varphi(x)$.

Définition 4.1. *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 $\mathbb{P}(.|\varphi)$ 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*

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

Définition 4.2 (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 4.3 (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*

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

5 Statistiques exhaustives minimales

Définition 5.1. *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 $\varphi = \xi \circ \varphi'$).*

Théorème 5.2 (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*

$$\varphi(x) = \varphi(x') \Longleftrightarrow \text{il existe } A(x, x') \in \mathbb{R} \text{ t.q. pour tout } \theta \in \Theta,\ f(x, \theta) = A(x, x')\, f(x', \theta)$$

alors φ est une statistique exhaustive minimale pour θ. Lors f ne s'annule pas, cela s'exprime plus simplement sous la forme

$$\varphi(x) = \varphi(x') \Longleftrightarrow \theta \mapsto \frac{f(x, \theta)}{f(x', \theta)} \text{ est une fonction indépendante de } \theta.$$

Définition 5.3. *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 5.4. *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 $g \circ \varphi \in F$, on a l'implication*

$$\left(\forall\theta \in \Theta : \int_{\mathbb{X}} g \circ \varphi d\mathbb{P}_\theta = 0\right) \Longrightarrow (\forall\theta \in \Theta : g \circ \varphi = 0 \quad \mathbb{P}_\theta - ps)$$

soit encore

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

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

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

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 5.6. *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 5.7 (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.*

6 Estimation ponctuelle

Définition 6.1 (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 6.2 (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 6.3 (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 6.4 (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 6.5 (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*

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

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.

7 Qualité d'un estimateur

Définition 7.1 (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 7.2 (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 7.3 (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 7.4. *On dit que $\widehat{\theta}_n$ est un estimateur consistant de θ si $\widehat{\theta}_n \to \theta$ en probabilité.*

On dit que $\widehat{\theta}_n$ est un estimateur fortement consistant de θ si $\widehat{\theta}_n \to \theta$ presque sûrement.

Finalement, on considérera que le meilleur estimateur possible de θ est un **estimateur sans biais et de variance minimale (ESBVM)**. Un tel estimateur n'existe pas forcément.

Théorème 7.5 (Rao-Blackwell)**.** *Soit $(E, \mathcal{E}, \mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\})$ un modèle paramétrique et $\underline{X} = (X_1, ..., X_n)$ un échantillon dans ce modèle. Soit $T(\underline{X})$ un estimateur de $g(\theta)$ de carré*

intégrable.
Si le modèle possède une statistique exhaustive $S(\underline{X})$ pour le paramètre θ, alors l'estimateur

$$\widetilde{T}(\underline{X}) = \mathbb{E}(T(\underline{X})|S(\underline{X}))$$

de $g(\theta)$ a un risque quadratique inférieur à $T(\underline{X})$ pour tout $\theta \in \Theta$.
Si $T(\underline{X})$ est un estimateur sans biais de $g(\theta)$ alors $\widetilde{T}(\underline{X})$ est également sans biais pour $g(\theta)$ et l'inégalité sur les risques quadratiques se traduit sur les variances.

Théorème 7.6 (Lehmann-Scheffé)**.** *Soit $(E, \mathcal{E}, \mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\})$ un modèle paramétrique et $\underline{X} = (X_1, ..., X_n)$ un échantillon dans ce modèle. Soit $T(\underline{X})$ un estimateur sans biais de $g(\theta)$ de carré intégrable et $S(\underline{X})$ une statistique exhaustive et complète de θ. Alors l'estimateur amélioré de Rao-Blackwell $\widetilde{T}(\underline{X}) = \mathbb{E}(T(\underline{X})|S(\underline{X}))$ est optimal dans la classe des estimateurs sans biais de $g(\theta)$.*

Définition 7.7 (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 7.8 (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) $\log L$ 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 pour lesquelles ces intégrales ont un sens,

$$\frac{\partial}{\partial \theta_j} \int_{\mathbb{X}} g(x)\mathcal{L}(x, \theta)d\mu(x) = \int_{\mathbb{X}} g(x)\frac{\partial \mathcal{L}}{\partial \theta_j}(x, \theta)d\mu(x) \quad et \tag{7.1}$$

$$\frac{\partial^2}{\partial \theta_j \partial \theta_k} \int_{\mathbb{X}} g(x)\mathcal{L}(x, \theta)d\mu(x) = \int_{\mathbb{X}} g(x)\frac{\partial^2 \mathcal{L}}{\partial \theta_j \partial \theta_k}(x, \theta)d\mu(x). \tag{7.2}$$

Une condition suffisante pour permuter la dérivée et l'intégrale est que pour tout $\theta \in \Theta$, il existe une boule B_θ centrée en θ telle que pour tous j, k,

$$\int_{\mathbb{X}} |g(x)| \sup_{\theta' \in B_\theta} \left| \frac{\partial \mathcal{L}}{\partial \theta'_j}(x, \theta') \right| d\mu(x) < +\infty \tag{7.3}$$

Définition 7.9 (Score)**.** *Sous (H1), (H2), (H3), le score pour la variable générique X est défini comme le vecteur gradient de $\log L(X, \theta)$ (i.e., la dérivée de $\log L(X, \theta)$ par rapport à θ) :*

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

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

Définition 7.10 (Quantité d'information de Fisher – cas multidimensionnel)**.** *Pour un échantillon $X_1, \dots, X_n$ et sous (H1), (H2), (H3), la matrice d'information de Fisher $I_n(\theta)$ est donnée par*

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

Définition 7.11 (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 7.12 (Simplification – cas multidimensionnel)**.** *Sous (H4), la quantité d'information peut se réécrire*

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

De plus l'information de Fisher est additive, ainsi

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

Proposition 7.13 (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 7.14 (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 7.15 (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

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

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

Définition 7.16 (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 7.17 (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* $\mathrm{Eff}\,(T_n) \to 1$*).*

Proposition 7.18 (propriété MME)**.** *Si* $\theta = \mathbb{E}\,[X]$ *et* X *a un moment d'ordre* 2 *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* θ*.*

Proposition 7.19 (Variance empirique)**.** *On suppose que* X *admet un moment d'ordre* 4*. La variance empirique* $S_n^2 = 1/n\sum_i (X_i - \overline{X}_n)^2$ *est un estimateur asymptotiquement sans biais de* $\mathrm{Var}\,(X)$*. La variance empirique corrigée*

$$S_{n,c}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

est un estimateur sans biais et convergent en moyenne quadratique de $\mathrm{Var}\,(X)$*.*

Proposition 7.20 (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*

- $\widehat{\theta}_n$ *converge presque sûrement vers* θ*.*
- $\sqrt{I_n(\theta)}(\widehat{\theta}_n - \theta)$ *converge en loi vers la loi normale* $\mathcal{N}(0,1)$*. C'est à dire* $\widehat{\theta}_n$ *est asymptotiquement gaussien* $\mathcal{N}(\theta, 1/(I_n(\theta)))$*.*

Proposition 7.21 (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,* $\varphi(\widehat{\theta}_n)$ *est asymptotiquement gaussien* $\mathcal{N}(\varphi(\theta), \varphi'(\theta)^2/(I_n(\theta)))$ *si* $\varphi'(\theta) \neq 0$*.*

Proposition 7.22 (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*

- $\widehat{\theta}_n$ *converge presque sûrement vers* θ.
- $\widehat{\theta}_n$ *est asymptotiquement gaussien* $\mathcal{N}(\theta, I_n(\theta)^{-1})$.

Proposition 7.23 (Méthode delta)**.** *Soit* $\varphi : \mathbb{R}^d \mapsto \mathbb{R}^p$ *est une fonction différentiable, dont on note* $J\varphi$ *sa matrice jacobienne. Sous H1 à H4, si* $\widehat{\theta}_n$ *est l'estimateur MLE de* $\theta \in \mathbb{R}^d$ *alors* $\varphi(\widehat{\theta}_n)$ *est l'estimateur MLE de* $\varphi(\theta)$*. De plus,* $\varphi(\widehat{\theta}_n)$ *est asymptotiquement gaussien* $\mathcal{N}_p(\varphi(\theta), J\varphi(\theta)I_n(\theta)^{-1}J\varphi(\theta)^T)$*, où* $I_n(\theta)$ *est la matrice d'information de Fisher (pour un échantillon de taille* n*).*

8 Régression linéaire

Définition 8.1. *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 8.2. *Le modèle de régression linéaire simple ou modèle linéaire simple est défini par*

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

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*.*

Si X et Y ne sont pas indépendantes, la connaissance de la valeur prise par X diminue notre incertitude sur la réalisation de Y. En effet, le théorème de la variance totale dit que

$$\mathrm{Var}\,(Y) = \mathbb{E}\left[\mathrm{Var}\,(Y|X)\right] + \mathrm{Var}\left(\mathbb{E}\left[Y|X\right]\right) \geqslant \mathbb{E}\left[\mathrm{Var}\,(Y|X)\right].$$

On cherche une fonction f telle que $\widehat{Y} = f(X)$ soit "le plus proche possible" de Y. Notons qu'il s'agit ici de prévision d'une variable, et non d'estimation d'un paramètre.

Quand on estime un paramètre, on souhaite que l'estimateur soit sans biais et de faible variance. De manière analogue, on cherchera une prévision $\widehat{Y}$ tel que :

$$\begin{aligned} \mathbb{E}\left[Y - \widehat{Y}\right] &= 0 \quad \text{prévision "sans biais"} \\ \mathrm{Var}\left(Y - \widehat{Y}\right) & \quad \text{soit la plus petite possible : faible variance de l'erreur de prévision.} \end{aligned}$$

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 $\prec X, Y \succ = \mathbb{E}[XY]$ 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 8.3. *On a*

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

L'intérêt des espaces de Hilbert est que l'on peut y définir la projection orthogonale sur un sous-espace vectoriel fermé. Par exemple, l'ensemble des variables constantes est le sous espace vectoriel de L^2 engendré par le vecteur $\mathbf{1}$. C'est donc une droite D dans L^2, et c'est un sous espace vectoriel fermé.

Théorème 8.4. *La projection orthogonale de X sur D 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

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

Théorème 8.5. *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 8.6. *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 8.7. *$\widehat{\beta}_1$ et $\widehat{\beta}_0$ sont des ESB de β_1 et β_0.*

Théorème 8.8. (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, $\forall i$, $\sigma^2 = \mathrm{Var}(\varepsilon_i) = \mathrm{Var}(Y_i - \beta_1 x_i - \beta_0)$, il est naturel d'estimer σ^2 par la variance empirique des $Y_i - \widehat{\beta}_1 x_i - \widehat{\beta}_0$.

Définition 8.9. *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 8.10. *$\widehat{\sigma}^2 = \frac{n}{n-2} S^2_{Y|x}$ 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 8.11. *Le modèle de régression linéaire simple gaussien est défini par*

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

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

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

Théorème 8.13. *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 $\frac{n-2}{n}\widehat{\sigma}^2$.*

En utilisant ces propriétés, on montre facilement que

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

Proposition 8.14 (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 8.15. *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 "$\sigma \neq \sigma_0$"*

$$W = \left\{\frac{(n-2)\widehat{\sigma}^2}{\sigma_0^2} < z_{n-2,1-\frac{\alpha}{2}} \quad ou \quad \frac{(n-2)\widehat{\sigma}^2}{\sigma_0^2} > z_{n-2,\frac{\alpha}{2}}\right\}.$$