



Statistiques et modélisation

Elena Di Bernardino

L3 MASS - 2021



Table des matières

I	Modélisation statistique	
1	Outils de probabilités	7
1.1	Loi d'une variable aléatoire réelle	7
1.1.1	Variables discrètes	7
1.1.2	Variables de loi absolument continue	7
1.1.3	Formules d'intégration	8
1.2	Paramètres de position	8
1.2.1	Espérance et variance	9
1.2.2	Coefficients d'asymétrie et d'aplatissement	9
1.2.3	Quantiles	10
1.3	Vecteurs gaussiens	10
1.3.1	Loi normale multivariée	10
1.3.2	Dérivées des lois gaussiennes	11
1.4	Convergence et théorèmes limites	12
1.4.1	Modes de convergence aléatoire	12
1.4.2	Théorème central limite	13
1.4.3	Travail en dimension d	14
II	Méthode d'estimation	
2	Échantillonnage et f° de répartition empirique	17
2.1	Situation et notations préliminaires	17

2.2	Estimation ponctuelle	17
2.2.1	Fonction de répartition empirique (ecdf)	17
2.2.2	Précision de l'estimation	18
2.2.3	Précision de l'estimation asymptotique	19
2.2.4	Précision non asymptotique	20
2.2.5	Décision	21
2.3	Estimation uniforme	22
2.3.1	Estimation uniforme	22
2.3.2	Intervalle de confiance uniforme	23
2.4	Estimation de fonctionnelles - Méthode de substitution	23
3	Méthode d'estimation en densité	27
3.1	Introduction	27
3.1.1	Notations et hypothèses	27
3.1.2	Familles paramétriques classiques	27
3.2	Méthode des moments	28
3.2.1	En dimension 1	28
3.2.2	En dimension d	28
3.3	Moments généralisés, M-estimateur et Z-estimateur	29
3.3.1	Z-estimateur	29
3.3.2	M-estimateur	29
3.3.3	Convergence de M-estimateur et Z-estimateur	30
3.3.4	Loi limite des Z-estimateurs et M-estimateurs	30
3.4	Maximum de vraisemblance	31
3.4.1	Principe du maximum de vraisemblance	31
3.4.2	Principe de vraisemblance pour 2 points	31
3.4.3	Passage à une famille de lois quelconque	32
3.4.4	Maximum de vraisemblance et M-estimateur	32
4	Méthode d'estimation en régression	35
4.1	Régression linéaire simple	35
4.1.1	Droite de régression	35
4.1.2	Moindres carrés et maximum de vraisemblance	36
4.1.3	Coefficient de détermination (R^2)	37
4.2	Régression linéaire multiple	38
4.2.1	Moindres carrés	38
4.2.2	Loi des estimateurs	39
4.2.3	Région de test critique	39



Modélisation statistique

1	Outils de probabilités	7
1.1	Loi d'une variable aléatoire réelle	
1.2	Paramètres de position	
1.3	Vecteurs gaussiens	
1.4	Convergence et théorèmes limites	

1. Outils de probabilités

1.1 Loi d'une variable aléatoire réelle

Soit E un espace probabilisé $\{\Omega, \mathcal{A}, \mathbb{P}\}$. Les probabilités $\omega \in \Omega$ sont des modélisations de l'univers des événements Ω .

1.1.1 Variables discrètes

Définition 1.1.1 Une variable aléatoire sur E est une application $x : (\Omega, \mathcal{A}) \rightarrow (\mathbb{R}, \mathcal{B})$, où \mathcal{B} est une tribu borélienne.

Définition 1.1.2 Une tribu borélienne est la plus petite tribu sur X contenant tous les ensembles ouverts.

Définition 1.1.3 La fonction de répartition de la variable réelle X est l'application $F : \mathbb{R} \rightarrow [0; 1]$ définie par :

$$F(x) = \mathbb{P}(X \leq x) = \mathbb{P}(\omega \in \Omega : X(\omega) \leq x) \quad (1.1)$$

1.1.2 Variables de loi absolument continue

Propriété 1.1.1 La fonction de répartition est croissante, continue à droite, tend vers 0 pour $x \rightarrow -\infty$ et 1 vers $+\infty$.

Définition 1.1.4 On note \mathbb{P}^* la loi d'une v.a., image de \mathbb{P} par X sur $(\mathbb{R}, \mathcal{B})$.

$$\mathbb{P}^*(A) = \mathbb{P}(x \in \mathcal{A}), \mathcal{A} \in \mathcal{B}(\mathbb{R}) \quad (1.2)$$

Définition 1.1.5 — Variable discrète. Une variable aléatoire dans \mathbb{R} , notée X est discrète si elle prend un ensemble de valeurs ou plan dénombrable $\{x_i, i \in \mathbb{N}\} \subset \mathbb{R}$.

Il suffit d'avoir les données des $\{(x_i, \mathbb{P}[X_i = x_i]), i \in \mathbb{N}\}$ pour déterminer intégralement la loi F .

- R** Si les points x_i sont isolés (X est à valeurs dans \mathbb{N} ou \mathbb{Z}), alors la f° de répartition de X est constante par morceaux, et les points de discontinuité de F sont les probabilités x_i .

Définition 1.1.6 La mesure de Lebesgue est une généralisation de la notion de volume. Cette mesure évolue dans les ensembles L^p .

Définition 1.1.7 — Fonctions à densité. Une v.a. dans \mathbb{R} , X est alors continue (à densité) si la f° de rép. s'écrit $F(x) = \int_{-\infty}^x f(t)dt, \forall x \in \mathbb{R}$, où dt est la mesure de Lebesgue sur \mathbb{R} . La fonction f est une densité de \mathbb{P} . Elle doit être :

- Positive ($f \geq 0$)
- Normalisée ($\int_{\mathbb{R}} f = 1$).

1.1.3 Formules d'intégration

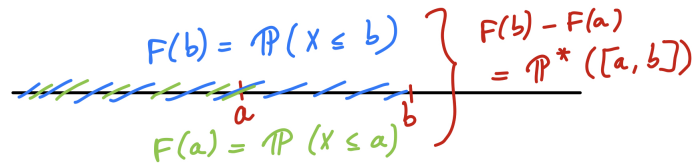
Propriété 1.1.2 Soit X une v.a. à valeurs dans \mathbb{R} , munie d'une loi F . Pour toute fonction test f , alors $\mathbb{E}[\varphi(X)] = \int_{\Omega} \varphi(X(\omega))\mathbb{P}(d\omega) = \int_{\mathbb{R}} \varphi(x)\mathbb{P}^*(dx)$. On écrit aussi :

$$\int_{\mathbb{R}} \varphi(x)\mathbb{P}^*(dx) = \int_{\mathbb{R}} \varphi(x)dF(x) = \mathbb{E}(\varphi(x)) \quad (1.3)$$

Ainsi, on a $F[X^2] = \int_{\mathbb{R}} x^2 dF(x)$.

- R** La propriété présente permet ainsi de dire qu'après modification des valeurs de la loi, il y a **conservation de la loi** initiale.

- R** Si on souhaite calculer $\mathbb{P}^*([a, b]) = F(b) - F(a)$.



Ainsi, ces f° de rép. permettent de calculer une probabilité sur des intervalles non bornés.

Théorème 1.1.3 — Théorème de transfert. Soit X une v.a. à valeurs dans $\{x_i, i \in \mathbb{N}\}$. Ainsi,

$$\mathbb{E}[\varphi(X)] = \int_{\mathbb{R}} \varphi(x)dF(x) = \sum_{i \in \mathbb{N}} \varphi(x_i)\mathbb{P}(X = x_i) \quad (1.4)$$

1.2 Paramètres de position

Une loi possède plusieurs moments (paramètres de la loi), comme la moyenne, la variance, le coefficient d'asymétrie (skewness), le coefficient d'acuité (kurtosis), les quantiles etc.

1.2.1 Espérance et variance

Définition 1.2.1 — Moment. Une v.a. X admet un moment d'ordre 2 lorsque :

$$\mathbb{E}[|X^2|] = \int |X(\omega)|^2 \mathbb{P}(d\omega) < +\infty \quad (1.5)$$

Dans ce cas, un moment d'ordre $p \in \mathbb{N} \setminus \{0\}$ est :

$$\mathbb{E}[X^p] = \int [X(\omega)]^p \mathbb{P}(d\omega) \quad (1.6)$$

Proposition 1.2.1 La moyenne μ_x existe : $u = F(X) = \int_{\Omega} X(\omega) \mathbb{P}(d\omega)$.

Définition 1.2.2 $\sigma_x^2 = \text{Var}(X)$ est le moment d'ordre 2 recentré de X :

$$\sigma_x = \mathbb{E}[X - \mu_x]^2 = \int_{\mathbb{R}} (x - \mu_x)^2 dF_x(x) = \sqrt{\text{Var}(X)} = \sqrt{\mathbb{E}[(x - \mu_x)^2]} \quad (1.7)$$

On calcule (s'ils existent !) tous les moments d'une loi :

Proposition 1.2.2 — Moment d'ordre p .

$$\mathbb{E}[X^p] = \int_{\mathbb{R}} x^p dF(x) = \begin{cases} \sum_{i \in \mathbb{N}} x_i^p \mathbb{P}(X = x_i) & \text{si } X \text{ est discrète.} \\ \int_{\mathbb{R}} x^p f(x) dx & \text{si } X \text{ est abs. } C^0. \end{cases} \quad (1.8)$$

1.2.2 Coefficients d'asymétrie et d'aplatissement

Définition 1.2.3 — Coef. d'asymétrie. La loi de X est symétrique par rapport à $\mu \in \mathbb{R}$ si $\forall x \in \mathbb{R}, F(\mu + x) = 1 - F(\mu - x)$, où F est la f° de rép. de X .

Dans le cas d'une loi à densité, si X est symétrique, alors en dérivant l'expression précédente, $\forall x, f(\mu + x) = f(\mu - x)$.

■ **Exemple** La fonction gaussienne centrée en 0. ■

Définition 1.2.4 Le coef. d'asymétrie de la variable aléatoire X (s'il existe un moment d'ordre 3) est défini par :

$$\alpha[X] = \frac{\mathbb{E}[(X - \mathbb{E}[X])^3]}{\sigma_X^3} \quad (1.9)$$

Attention, ce coefficient $\alpha(x)$ est une mesure **grossière** de la symétrie.

- Si X est symétrique, alors $\alpha(x) = 0$.
- Si $\alpha(x) = 0$, alors X n'est pas forcément symétrique (i.e. la réciproque est **fausse** !)

Définition 1.2.5 — Aplatissement. Le coef. de symétrie de la variable aléatoire X (s'il existe un moment d'ordre 4) est défini par :

$$K[X] = \frac{\mathbb{E}[(X - \mathbb{E}[X])^4]}{\sigma_X^4} \quad (1.10)$$

- Si $X \sim N(0, 1)$, alors $K(X) = 0$.
- Si $K(X) < 0$, alors les piques de distribution de X sont plus légères que la loi normale.
- Sinon, elles sont plus lourdes.

R Graphiquement, si $K(X) < 0$, la courbe est plus aplatie en son pic ; sinon, elle est plus pointue.

1.2.3 Quantiles

Définition 1.2.6 Si X est une v.a. dans \mathbb{R} une f° de rép. F , alors la quantité d'ordre p , $p \in (0, 1)$ est définie comme l'**unique** solution q_i du problème :

$$F_X(q_p) = p \Leftrightarrow \mathbb{P}(X = q_p) = p \Leftrightarrow F_X^{-1}(p) = q_p \quad (1.11)$$

■ **Exemple** Quelle est la probabilité qu'une vague frappant le port de Nice soit de hauteur $5m$, avec une probabilité d'échec de 99% ? ■

Définition 1.2.7 Le quantile d'ordre p de X est la quantité :

$$q_p = \frac{1}{2} [\inf\{x, F(x) > p\} + \sup\{I(x) < p\}] \quad (1.12)$$

Concrètement, si on divise le jeu de données en n paquets, il s'agit du p -ième.

Propriété 1.2.3 La médiane de X de la loi F_X est par définition $q_{\frac{1}{2}}$. Elle respecte :

- $\mathbb{P}(X \geq \frac{1}{2}) \geq \frac{1}{2}$
- $\mathbb{P}(X \leq \frac{1}{2}) \geq \frac{1}{2}$

1.3 Vecteurs gaussiens

1.3.1 Loi normale multivariée

Définition 1.3.1 — Vect. gaussien. Si $X = {}^t(X_1, \dots, X_n)$ est un vecteur aléatoire en \mathbb{R}^n , son espérance est **si elle existe** :

$$\mathbb{E}[X] = (\mathbb{E}[X_1], \dots, \mathbb{E}[X_n]) \quad (1.13)$$

La variance de X est la matrice $\Sigma_X = \mathbb{E}[(X - \mathbb{E}[X])^t(X - \mathbb{E}[X])]$, appelée matrice de variance covariance. Ce vecteur Σ existe **que** si $\mathbb{E}[||X||] < +\infty$.

■ **Exemple** Si $X = (X_1, X_2)$, alors $\Sigma_X = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) \\ \text{Cov}(X_1, X_2) & \text{Var}(X_2) \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$. ■

Propriété 1.3.1 — $n > 1$.

- $\Sigma_X = \mathbb{E}[{}^tXX] - \mathbb{E}[X]\mathbb{E}[{}^tX]$
- $\forall a \in \mathbb{R}^n, \text{Var}({}^taX) = {}^ta\Sigma_X a$
- Si $b \in \mathbb{R}^k, A \in M_{k \times n}$, on a $\Sigma_{AX+b} = A\Sigma_X A^t$

■ **Rappel** En dimension $n = 1$:

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

Définition 1.3.2 — Vecteurs Gaussiens. On note $X \sim N_1(a, I_n)$ un vecteur $X = {}^t(\xi_1, \dots, \xi_n)$ dont ses lois marginales sont des lois $N(0, 1)$, indépendantes entre elles.

Propriété 1.3.2 (X_1, \dots, X_n) sont indépendantes $\Leftrightarrow \mathbb{P}(X_1 \leq x_1, \dots, x_n \leq x_n) = \sum_{i=1}^n \mathbb{P}(X_i \leq x_i)$

Propriété 1.3.3 La moyenne de $X \sim N(a, I_n)$ est nulle et sa matrice de variance-covariance est l'identité. La loi X est alors continue de densité par rapport à la mesure de Lebesgue sur \mathbb{R}^n .

■ **Rappel** Dans le cas $n = 1$, si $Z \sim N(\mu, \sigma^2)$, alors $T = \frac{Z - \mu}{\sigma} \sim N(0, 1)$

Définition 1.3.3 Un vecteur aléatoire X à valeurs dans \mathbb{R}^n , de forme :

$$\forall n \in \mathbb{N}, f_X(x) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2} {}^t x x\right) \quad (1.14)$$

Il est Gaussien (normal) si, pour $A \in M_n$ et un vecteur $\mu \in \mathbb{R}^n$, on a :

$$X = \mu + A\xi, \xi \sim N(a, I_n) \quad (1.15)$$

Propriété 1.3.4 Un vecteur aléatoire X est dit Gaussien si et seulement si toute combinaison linéaire des composantes de X est une variable aléatoire à valeurs dans \mathbb{R} .

Dans le cas $n = 2$:

$$(X_1, X_2) \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right) \Rightarrow \forall (a, b) \in \mathbb{R}^2 \begin{cases} X_1 + X_2 \sim N_1(\dots) \\ aX_1 + bX_2 \sim N_1(\dots) \end{cases} \quad (1.16)$$

Définition 1.3.4 — Densité. Soit $X \sim N_m(\mu, \Sigma_X)$. Alors :

$$f_X(x) = \frac{1}{(2\pi)^{n/2} \sqrt{\det(\Sigma_X)}} \exp\left(-\frac{1}{2} {}^t (X - \mu) \Sigma_X^{-1} (x - \mu)\right) = \sum_{i=1}^n \underbrace{f_{X_i}(x_i)}_{N_1(\mu_i, 1)} \quad (1.17)$$

1.3.2 Dérivées des lois gaussiennes

Ainsi, on va étudier trois lois (les mêmes qu'en économétrie) :

- ★ Loi de χ^2 .
- ★ Loi de Student
- ★ Loi de Fischer

Définition 1.3.5 — χ^2 . Soit Y une v.a. à valeurs dans \mathbb{R} . Alors :

$$Y = \sum_{i=1}^n X_i^2, X_i \sim N(0, 1) \quad (1.18)$$

Ces X_i sont des v.a. identiquement distribuées à valeurs dans \mathbb{R}_+ . On dit que $Y \sim \chi^2(x)$ i.e. si $X \sim N(0, I_n)$, alors $|X|^2 \sim \chi^2(n)$.

Définition 1.3.6 — Student. Soit une v.a. $T \sim t(n)$. Si $T := \frac{\xi}{\sqrt{Y/m}}$, où $\xi \sim N(0, 1)$, $Y \sim \chi^2(n)$ et ξ indépendante de Y , à valeurs dans \mathbb{R} , on dit que cette loi est de Student.

R La loi $t(m)$ est plus "dispersée" que la loi normale sur $[0, 1]$. Si $T \sim t(m)$ et que $\xi \sim N(0, 1)$, on a par exemple $K[T] > K[\xi]$. Dans le cas extrême où $n = 1$, K n'est même plus définie.

Définition 1.3.7 Soit une v.a. $Y \sim \text{Fisher}(p, q)$, admettant p, q degrés de liberté.

Si $Y = \frac{U}{V}$ où $U \sim \chi^2(p)$ et $V \sim \chi^2(q)$, on dit que cette loi est de Fischer.

1.4 Convergence et théorèmes limites

1.4.1 Modes de convergence aléatoire

R Dans cette section, toutes les limites, si cela n'est pas précisé, est en $+\infty$.

On considère une suite (ξ_n) de v.a. réelle, dans $(\Omega, \mathcal{A}, \mathbb{P})$.

■ **Exemple** Soit (X_1, \dots, X_n) des v.a. réelle, alors $\bar{X}_n = \frac{1}{n} \sum X_i$ correspond à la définition. ■

Définition 1.4.1 — Convergence en probabilité. La suite (ξ_n) converge vers ξ en probabilité (notée $\xi_n \xrightarrow{\mathbb{P}} \xi$ si, $\forall \varepsilon > 0$, on ait :

$$\lim_{n \rightarrow +\infty} \mathbb{P}[|\xi_n - \xi| \geq \varepsilon] = 0 \quad (1.19)$$

Définition 1.4.2 — Convergence presque sûre. La suite ξ_n converge presque sûrement (notée $\xi_n \xrightarrow{\text{p.s.}} \xi$) si :

$$\mathbb{P}[\limsup |\xi_n - \xi| > 0] = 0 \quad (1.20)$$

Définition 1.4.3 — Convergence L^p (moments). À $p \in [0, +\infty[$ fixé, la suite η_n converge dans L^p si :

$$\lim_{n \rightarrow +\infty} \mathbb{E}[|\xi_n - \xi|^p] = 0 \quad (1.21)$$

R Les convergences p.s. et L^p entraînent la convergence en probabilité, mais la **réciproque** est vraie que si l'on ajoute des conditions supplémentaires :

- ★ Pour L^p , si $\xi_n \xrightarrow{\mathbb{P}} \xi$, que $n|\xi_n| < \eta$ et que $\mathbb{E}[\eta^p] < +\infty$ pour un $p > 0$, alors la réciproque inverse est vraie.
- ★ Pour celle p.s., si f est C^0 et $\xi_n \xrightarrow{\mathbb{P}} \xi$, alors $f(\xi_n) \xrightarrow{\mathbb{P}} f(\xi)$. (*continuous mapping theorem*).

De plus, les convergences p.s. et L^p ne peuvent être entraînées entre elles.

Elle traduit la propriété statistique suivante :

Propriété 1.4.1 Peu importe le niveau de risque $\alpha > 0$ et la précision $\varepsilon > 0$, il existe un rang $n_{(\varepsilon, \alpha)}$ à partir duquel on peut "affirmer" que ξ_n approche ξ avec une erreur inférieure à ε . La probabilité pour que cette affirmation soit fausse est inférieure à α . Pour $n_{(\varepsilon, \alpha)}$, $\mathbb{P}(|\xi_n - \xi| \leq \varepsilon) \geq 1 - \alpha$.

Définition 1.4.4 — Convergence en loi. La suite ξ_n converge en loi vers ξ si, pour toute fonction test φ continue et bornée, on ait :

$$\mathbb{E}[\varphi(\xi_n)] \xrightarrow{n \rightarrow +\infty} \mathbb{E}[\varphi(\xi)] \quad (1.22)$$

R On peut remplacer ξ_n par un vecteur aléatoire ε_n à valeurs dans \mathbb{R}^k . La convergence en loi est plus faible que les autres convergences, mais elle est facilement adaptable à une dimension $k \geq 2$.

1.4.2 Théorème central limite

Définition 1.4.5 — Convergence en distribution. La suite $\xi_n \rightarrow \xi$ converge en distribution si et seulement si $\forall \varphi$ fonction continue et bornée, on a :

$$\mathbb{E}[\varphi(\xi_n)] \xrightarrow{d} \mathbb{E}[\varphi(\xi)] \quad (1.23)$$

Propriété 1.4.2 1. La suite de vecteurs aléatoires ξ_n à valeurs dans \mathbb{R}^d converge en loi vers ξ .
2. Si $\xi_n \rightarrow \xi$ en loi, que $g : \mathbb{R} \rightarrow \mathbb{R}$, alors $\forall a \in \mathbb{R}^d$, si $t a \xi_n \rightarrow \xi$ continue, alors $g(\xi_n) \rightarrow g(\xi)$.

Propriété 1.4.3 — Slutsky. Si $\xi_n \xrightarrow{d} \xi$ et $\eta_n \xrightarrow{\mathbb{P}} \eta$, alors $(\xi_n, \eta_n) \xrightarrow{d} (\xi, \eta)$. En particulier, si $h : \mathbb{R}^2 \rightarrow \mathbb{R}$ est continue, alors $h(\xi_n, \eta_n) \xrightarrow{d} h(\xi, \eta)$.

Propriété 1.4.4 Quelques propriétés :

- Si $h(x+y) = x+y$ et que $h(xy) = xy$, alors $\xi_n + \eta_n \rightarrow \xi + \eta$ et $\xi_n * \eta_n \rightarrow \xi * \eta$.
- Si X_1, \dots, X_n une suite de v.a. à valeurs dans \mathbb{R} , on notera $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$.
- Si X_1, \dots, X_n sont indépendantes et de même loi F , on notera $X_1, \dots, X_n \underset{\text{ind}}{\sim} F$.

Propriété 1.4.5 Soient $X_1, \dots, X_n \underset{\text{ind}}{\sim} F$ tels que $\text{Var}(X_i) = \sigma^2 < +\infty$. on note $\mu = \mathbb{E}[X]$. Alors :

$$\mathbb{E}[\bar{X}_n] = \mu \text{ et } \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n} \quad (1.24)$$

R La propriété ci-dessus (1.4.5) implique que $\bar{X}_n \xrightarrow{L^2} \mu$ et donc que $\bar{X}_n \xrightarrow{\mathbb{P}} \mu$.

Théorème 1.4.6 — Convergence p.s. de \bar{X}_n . Soient $X_1, \dots, X_n \underset{\text{ind}}{\sim} F$. Alors si $\mathbb{E}[|X|] < +\infty$, en notant $\mu = \mathbb{E}[X]$, on a $\bar{X}_n \xrightarrow{p.s.} \mu$.

R
$$\mathbb{E}[(\sqrt{n}(\bar{X}_n - \mathbb{E}(\bar{X}_n)))^2] = n\mathbb{E}[(\bar{X}_n - \mathbb{E}[\bar{X}_n])^2] = n * \text{Var}(\bar{X}_n) = n \frac{\sigma^2}{n} = \sigma^2 \quad (1.25)$$

On cherche le comportement asymptotique de $\sqrt{n}(\bar{X}_n - \mu)$.

Théorème 1.4.7 — Théorème central limite. Soient $X_1, \dots, X_n \underset{\text{ind}}{\sim} F$ tels que $\mathbb{E}[X^2] < +\infty$ et $\sigma^2 = \text{Var}(X) > 0$. Alors :

$$\sqrt{n} \left(\frac{\bar{X}_n - \mu}{\sigma} \right) \xrightarrow{d} N(0, 1) = Y \quad (1.26)$$

R On note en général que ξ_n est asymptotiquement normale si $\sqrt{n}(\xi_n - \xi) \xrightarrow{d} N(0, \sigma^2)$.

Propriété 1.4.8 — Δ -méthode. Si ξ_n est asymptotiquement normale et qu'on a $g : \mathbb{R} \rightarrow \mathbb{R}$ une fonction $C^1(\mathbb{R})$, alors $g(\xi_n)$ est également asymptotiquement normale. De plus :

$$\sqrt{n}(g(\xi_n) - g(\xi)) \xrightarrow{d} N(0, g'(\xi)^2 \sigma^2) \quad (1.27)$$

Démonstration. Soit la fonction $h(x) = \begin{cases} \frac{g(x) - g(\mu)}{x - \mu} & \text{si } x \neq \mu \\ g'(\mu) & \text{sinon.} \end{cases}$

Comme ξ_n est asymptotiquement normale, alors $\xi_n \xrightarrow{\mathbb{P}} \mu$ et donc $g(\xi_n) \xrightarrow{\mathbb{P}} g(\mu)$ et $h(\xi_n) \xrightarrow{\mathbb{P}} h(\mu) =$

$g'(\mu)$. Ainsi, $\sqrt{n}(g(\xi_n) - g(\xi)) = h(\xi_n) * \eta_n$, en posant $\eta_n = \sqrt{n}(\xi_n - \mu)$.

$$\text{Ainsi, } \sqrt{n}(g(\xi_n) - g(\xi)) = \left(\frac{g(\xi_n) - g(\xi)}{\xi_n - \mu} \right) \sqrt{n}(\xi_n - \mu).$$

Donc d'après la propriété de Slutsky, $\sqrt{n}(g(\xi_n) - g(\mu)) \xrightarrow{d} g'(\mu)N(0, \sigma^2) = N(0, \sigma^2 g'(\mu)^2)$. ■

1.4.3 Travail en dimension d

Théorème 1.4.9 — Théorème Central Limite. Soient X_1, \dots, X_n une suite de vecteur aléatoires dans \mathbb{R}^d , i.i.d. tel que $\mathbb{E}[\|X\|_2] < +\infty$. En notant $\mu = \mathbb{E}[X]$, μ la matrice de covariance de X , on a :

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\text{dist}} N(0, \Sigma) \quad (1.28)$$

Propriété 1.4.10 — Δ -méthode, dim. d . Soient ξ_1, \dots, ξ_n n vecteurs aléatoires en \mathbb{R}^d , asymptotiquement normaux, tels que $\sqrt{n}(\xi_n - \mu) \rightarrow N_d(0, \Sigma_n)$, où $\mu \in \mathbb{R}^d$, Σ la matrice de covariance de ξ . Alors, si $g : \mathbb{R}^d \rightarrow \mathbb{R}^k$ une fonction différentiable, on a :

$$\sqrt{n}(g(\xi_n) - g(\mu)) \xrightarrow{d} N(0, \underbrace{J_g(\mu) \Sigma J_g(\mu)}_{\sim g'(\xi)^2 \sigma^2}) \quad (1.29)$$

où $J_g(x) = \begin{pmatrix} \partial_1 g_1(x) & \cdots & \partial_d g_1(x) \\ \vdots & \ddots & \vdots \\ \partial_1 g_n(x) & \cdots & \partial_d g_n(x) \end{pmatrix}$ est la matrice Jacobienne de x .



Méthode d'estimation

2	Échantillonnage et f° de répartition empirique	17
2.1	Situation et notations préliminaires	
2.2	Estimation ponctuelle	
2.3	Estimation uniforme	
2.4	Estimation de fonctionnelles - Méthode de substitution	
3	Méthode d'estimation en densité	27
3.1	Introduction	
3.2	Méthode des moments	
3.3	Moments généralisés, M-estimateur et Z-estimateur	
3.4	Maximum de vraisemblance	
4	Méthode d'estimation en régression	35
4.1	Régression linéaire simple	
4.2	Régression linéaire multiple	

2. Échantillonnage et f° de répartition empirique

2.1 Situation et notations préliminaires

On observe (X_1, \dots, X_n) de loi **inconnue** F , à valeurs dans \mathbb{R} . On cherche à estimer une fonctionnelle T de F , notée $T(F) \in \mathbb{R}$. On dénote notre estimateur \hat{T}_n , qui est une v.a. ne dépendant que de X_1, \dots, X_n et pas de F (qui est inconnue !). On écrit donc $\hat{T}_n = g_n(X_1, \dots, X_n)$ pour une certaine fonction $g_n : \mathbb{R}^n \rightarrow \mathbb{R}$, qui ne dépend pas de F .

2.2 Estimation ponctuelle

Dans tout ce qui suit, on suppose que tous les théorèmes fonctionnent en dimension d .

2.2.1 Fonction de répartition empirique (ecdf)

Définition 2.2.1 La fonction de répartition empirique s'écrit :

$$F(x) = \mathbb{P}(X \leq x) \quad (2.1)$$

Soit $x_0 \in \mathbb{R}$, et (X_1, \dots, X_n) observations. Que peut-on dire de $F(x_0) = P(X = x_0)$? L'idée la plus simple pour estimer $F(x_0)$ est d'introduire :

$$\frac{1}{n} \text{Card}\{X_i \in]-\infty; x_0], i = [1; n]\} \quad (2.2)$$

Définition 2.2.2 Une fonction de répartition empirique :

$$\hat{F}_n(x_0) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{X_i \leq x_0\}}, x_0 \in \mathbb{R} \quad (2.3)$$

Propriété 2.2.1

1. $\mathbb{E}[\hat{F}_n(x_0)] = F(x_0)$
2. $\text{Var}(\hat{F}_n(x_0)) = \mathbb{E}[(\hat{F}_n(x_0) - \mathbb{E}(\hat{F}_n(x_0)))^2] = \mathbb{E}[(\hat{F}_n(x_0) - F(x_0))^2] = \frac{F(x_0)(1 - F(x_0))}{n}$

3. Alors $\hat{F}_n(x_0) \xrightarrow{L^2} F(x_0)$ et donc $\hat{F}_n(x_0) \xrightarrow{\mathbb{P}} F(x_0)$.

Démonstration. L'indicatrice $\mathbf{1}_{\{X_i \leq x_0\}}$ suit une loi de Bernoulli, de paramètre $p = F(x_0)$. Alors $n\hat{F}_n(x_0) \sim B(n, p)$. L'espérance vaut $nF(x_0)$ et la variance vaut $nF(x_0)(1 - F(x_0))$. ■

Théorème 2.2.2 — Loi forte des grands nombres.

$$\hat{F}_n(x_0) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{X_i \leq x_0\}} \xrightarrow{p.s.} F(x_0) \quad (2.4)$$

2.2.2 Précision de l'estimation

Définition 2.2.3 Soit $l(x, y) = (x - y)^2$, x, y deux réels (perte quadratique). Alors :

$$\sup_{F \in \mathcal{F}} |\mathbb{E}[l(\hat{F}_n(x_0), F(x_0))]| = \sup_{F \in \mathcal{F}} \left(\frac{F(x_0)(1 - F(x_0))}{n} \right) = \frac{1}{4n} \quad (2.5)$$

R Grosso modo, la pire estimation possible sera en $\frac{1}{4n}$.

Une manière pour quantifier la précision est :

Propriété 2.2.3

$$\mathbb{P}(|\hat{F}_n(x_0) - F(x_0)| \geq t) \leq \frac{1}{t^2} \text{Var}(\hat{F}_n(x_0)) \leq \frac{1}{t^2 4n} \quad (2.6)$$

R La précision augmente avec le nombre de valeurs n , mais baisse si t est peu restrictif.

■ **Rappel — Markov.** $\forall t > 0, \mathbb{P}(|X - \mu_X| \geq t) \leq \frac{\sigma_X^2}{t^2}$

Proposition 2.2.4 Posons $\alpha \in [0, 1]$, et prenons un $t = t(\alpha, n)$ le plus petit possible, de sorte que $\frac{1}{4nt^2} \leq \alpha$. Ainsi, $\frac{1}{4nt^2} \leq \alpha \Rightarrow t = \frac{1}{2\sqrt{n\alpha}}$.

On en déduit donc l'intervalle de confiance suivant :

Propriété 2.2.5

$$I_{n,\alpha} = [\hat{F}_n(x_0) - \frac{1}{2\sqrt{n\alpha}}, \hat{F}_n(x_0) + \frac{1}{2\sqrt{n\alpha}}] \quad (2.7)$$

Cet intervalle contient $F(x_0)$, avec une probabilité $1 - \alpha$.

Définition 2.2.4 L'intervalle $I_{n,\alpha}$ est appelé intervalle de confiance pour la valeur inconnue $F(x_0)$, au niveau de risque α . La propriété $\mathbb{P}(F(x_0) \in I_{n,\alpha}) \geq 1 - \alpha$ est dite de couverture.

R On applique l'inégalité de Markov à n fixé, ce qui donne :

$$|I_{n,\alpha}| = \frac{1}{2\sqrt{n\alpha}} \begin{cases} \xrightarrow{n \rightarrow \infty} 0 \\ \xrightarrow{\alpha \rightarrow 0} +\infty \end{cases} \quad (2.8)$$

2.2.3 Précision de l'estimation asymptotique

On voudrait considérer $\sqrt{n}(\hat{F}_n(x_0) - F(x_0))$. On a :

Propriété 2.2.6 — TCL pour la ecdf.

$$\xi_n = \sqrt{n} \left(\frac{\hat{F}_n(x_0) - F(x_0)}{\sqrt{\hat{F}_n(x_0)(1 - \hat{F}_n(x_0))}} \right) \quad (2.9)$$

De plus,

$$\mathbb{P}(\xi_n \in [\pm \Phi^{-1}(1 - \frac{\alpha}{2})]) \rightarrow 1 - \alpha \quad (2.10)$$

R On note Φ la fonction de rép. de $N(0, 1)$ et Φ^{-1} son quantile.

Démonstration. $\sqrt{n}(\frac{\bar{X}_n - \mu}{\sigma}) \rightarrow N(0, 1)$: TCL pour la somme de v.a. i.i.d. On a $\hat{F}_n(x_0) = \frac{1}{n} \sum \mathbf{1}_{\{x_i \leq x_0\}}$. En remplaçant \bar{X}_n , on a :

$$\sqrt{n} \left(\frac{\hat{F}_n(x_0) - F(x_0)}{\sqrt{F_0(x)(1 - F_0(x))}} \right) \xrightarrow{d} N(0, 1)$$

On va multiplier et diviser par $\sqrt{\hat{F}_n(x_0)(1 - \hat{F}_n(x_0))}$, de probabilité égale à 1, ce qui ne changera pas la probabilité. Grâce au TCD, on conserve la convergence vers $N(0, 1)$. Ainsi :

$$\xi_n = \frac{\sqrt{n}(\hat{F}_n(x_0) - F(x_0))}{\sqrt{\hat{F}_n(x_0)(1 - \hat{F}_n(x_0))}} \xrightarrow{d} N(0, 1)$$

On a $\xi_n \xrightarrow{d} N(0, 1)$, donc :

$$\begin{aligned} \mathbb{P}[\xi_n \in [\pm \Phi^{-1}(1 - \frac{\alpha}{2})]] &\xrightarrow{d} \Phi(\Phi^{-1}(1 - \frac{\alpha}{2})) - (\Phi(-\Phi^{-1}(1 - \frac{\alpha}{2}))) \\ &= 1 - \frac{\alpha}{2} - \Phi(\Phi^{-1}(\frac{\alpha}{2})) = 1 - \alpha \end{aligned}$$

■

On peut donc interpréter le 2ème point de la prop. 17 comme $qnorm(1 - \alpha)$:

— Lorsque " n est grand", $\xi_n := \left(\frac{\hat{F}_n(x_0) - F(x_0)}{\sqrt{\hat{F}_n(x_0)(1 - \hat{F}_n(x_0))}} \right) \in [\pm \Phi^{-1}(1 - \frac{\alpha}{2})]$.

R On obtient le fameux intervalle $[-1, 96; 1, 96]$ pour l'intervalle de confiance à 95%.

On en déduit que $J_{n,\alpha} = [\hat{F}_n(x_0) \pm \sqrt{\frac{\hat{F}_n(x_0)(1 - \hat{F}_n(x_0))}{n}}] \Phi^{-1}(1 - \frac{\alpha}{2})$. Ainsi $\mathbb{P}(F(x_0) - J_{n,\alpha}) \rightarrow 1 - \alpha$. (prop. de couverture)

Propriété 2.2.7 La précision de $J_{n,\alpha} = \sqrt{\frac{\hat{F}_n(x_0)(1 - \hat{F}_n(x_0))}{n}}] \Phi^{-1}(1 - \frac{\alpha}{2})$:

- Comportement en n de l'ordre de $\frac{1}{\sqrt{n}}$
- $\frac{1}{\alpha} \rightarrow +\infty, \Phi^{-1}(1 - \frac{\alpha}{2}) \rightarrow +\infty$
- $\Phi^{-1}(1 - \frac{\alpha}{2}) \ll (\sqrt{\alpha})^{-1}$
- $\sqrt{\hat{F}_n(x_0)(1 - \hat{F}_n(x_0))} < \frac{1}{2}$

2.2.4 Précision non asymptotique

Propriété 2.2.8 — Intervalle de confiance non paramétrique de la f.d.r. $F(x_0)$ en x_0 . On a les deux intervalles suivants :

- Markov non-asymptotique : $I_{n,\alpha} = \frac{1}{2\sqrt{n\alpha}}$
- TCL asymptotique : $J_{n,\alpha} = \frac{2}{\sqrt{n}} \sqrt{\hat{F}_n(x_0)(1 - \hat{F}_n(x_0))} \Phi^{-1}(1 - \frac{\alpha}{2})$

Théorème 2.2.9 — Inégalité de Hoeffding. Soient Y_1, \dots, Y_n v.a. à valeurs dans \mathbb{R} tels que $\forall i, \mathbb{E}[Y_i] = 0$ et $a_i \leq Y_i \leq b_i$. Soit $t > 0$, alors $\forall \lambda > 0$, on a :

$$\mathbb{P}(\sum_{i=1}^n Y_i \geq t) \leq e^{-\lambda t} \prod_{i=1}^n \exp((\frac{\lambda^2(b_i - a_i)}{8})) \quad (2.11)$$

Corollaire 2.2.10 Si on a X_1, \dots, X_n n v.a. de Bernoulli, de même paramètre p , et si $\bar{X}_n = \frac{1}{n} \sum X_i$, alors $\forall t > 0$, on a :

$$\mathbb{P}(|\bar{X}_n - \mu| \geq t) \leq 2 \exp(-2nt^2) \quad (2.12)$$

Démonstration. On applique le théorème de Hoeffding avec $Y_i = X_i - p$ ($\mathbb{E}[Y_i] = 0$). Ainsi, $a_i \leq Y_i \leq a_i + 1$. Cela donne :

$$\mathbb{P}(\sum_{i=1}^n Y_i \geq t) \leq e^{-\lambda t} \prod_{i=1}^n \exp((\frac{\lambda^2(1)^2}{8}))$$

En prenant $\lambda = \frac{4t}{n}$, on obtient $\mathbb{P}(\sum Y_i \geq t) \leq \exp(-\frac{4t}{n}t) \exp(\frac{-\lambda^2}{8})^n = \exp(-\frac{2t^2}{n})$.

Cela revient à écrire :

$$\mathbb{P}(\bar{X}_n - p \geq t) = \mathbb{P}(\sum Y_i \geq nt) \leq \exp(-2nt^2)$$

De même :

$$\mathbb{P}(\bar{X}_n - p \leq -t) \leq \exp(-2nt^2)$$

Et ainsi :

$$\mathbb{P}(|\bar{X}_n - p| \geq t) = \mathbb{P}(\bar{X}_n - p \leq -t) + \mathbb{P}(\bar{X}_n - p \geq t) \leq 2 \exp(-2nt^2)$$

■

Propriété 2.2.11 Pour tout $\alpha > 0$, l'intervalle $I_{n,\alpha}$ ci-dessous est un intervalle de confiance pour $F(x_0)$, de niveau $1 - \alpha$.

$$I_{n,\alpha}^* = [\hat{F}_n(x_0) \pm \sqrt{\frac{1}{2n} \ln(\frac{2}{\alpha})}] \quad (2.13)$$

Démonstration. L'indicatrice $\mathbf{1}_{X_i \leq x_0}$ suit une loi de Bernoulli, de paramètre $F(x_0)$. On a donc $\forall t > 0$:

$$\mathbb{P}(|\hat{F}_n(x_0) - \mathbb{E}[\hat{F}_n(x_0)]| \geq t) = \mathbb{P}(|\hat{F}_n(x_0) - F(x_0)| \geq t) \leq e \exp(-2nt^2)$$

R Dans $F_n(x_0) = \frac{1}{n} \sum X_i$, on peut remplacer par $\frac{1}{n} \sum \mathbf{1}_{X_i \leq x_0}$.

On cherche $t = t(n, \alpha)$ le plus petit possible de sorte que $2\exp(-2nt^2) \leq \alpha$. On cherche à forcer le fait que $\exp(-2nt^2) = \frac{\alpha}{2}$.

$$\begin{aligned} \frac{\alpha}{2} \exp(-2nt^2) &\stackrel{!}{=} \frac{\alpha}{2} \Rightarrow \ln(\exp(-2nt^2)) = \ln\left(\frac{\alpha}{2}\right) \\ \Rightarrow 2nt^2 &= -\ln\left(\frac{\alpha}{2}\right) = \ln\left(\frac{2}{\alpha}\right) \Rightarrow t = t(n, \alpha) = \sqrt{\frac{1}{2n} \ln\left(\frac{2}{\alpha}\right)} \end{aligned}$$

■

R On peut comparer les deux intervalles $|I_{n,\alpha}|$ et $|I_{n,\alpha}^*|$:

$$\frac{|I_{n,\alpha}^*|}{|I_{n,\alpha}|} = \frac{\sqrt{\frac{1}{2n} \ln\left(\frac{2}{\alpha}\right)}}{\frac{1}{2\sqrt{n\alpha}}} = \frac{2}{\sqrt{2}} \sqrt{\alpha \ln\left(\frac{2}{\alpha}\right)} \xrightarrow{\alpha \rightarrow 0} 0 \quad (2.14)$$

Ainsi, les deux intervalles $|I_{n,\alpha}|$ et $|I_{n,\alpha}^*|$ convergent vers 0, mais $|I_{n,\alpha}^*|$ converge plus rapidement vers 0 lorsque $\alpha \rightarrow 0$.

■ **Exemple** Si on prend $\alpha = 1\%$, alors $\frac{|I_{n,\alpha}^*|}{|I_{n,\alpha}|} = 0.33$. Cela veut dire qu'on a une précision 3 fois plus élevée pour $|I_{n,\alpha}^*|$ que $|I_{n,\alpha}|$. ■

2.2.5 Décision

La problématique dans cette partie est de savoir si on peut **tester** cette loi, avec X un vecteur $\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \sim \hat{F}_n(x_0)$.

Soit F_0 une distribution donnée. On souhaite savoir, en vu de l'échantillon de taille n : (X_1, \dots, X_n) , de loi $F \in \mathcal{F}$, si $\boxed{F(x_0) = F_0(x_0)}$.

On formule le problème de la manière suivante : on construit à partir des observations une "procédure" (estimateur) $\varphi_n = \varphi_n(X_1, \dots, X_n) \in \{0, 1\}$.

R Comme $\varphi_n \in \{0, 1\}$, on dira que ce test est "simple".

On associe aux valeurs $\varphi_n = 0$ la réponse "oui" et à 1 la valeur "non" à la question précédente.

On dira donc qu'on teste l'hypothèse nulle H_0 vs. l'hypothèse alternative $H_1 : \{F(x_0) \neq F_0(x_0)\}$.

Si φ_n est un test simple, on peut le représenter comme une indicatrice $\mathbf{1}_{\{(x_1, \dots, x_n) \in R_n\}}$, avec $R_n \subset \mathbb{R}^n$ un sous-ensemble de l'espace des observations.

■ **Définition 2.2.5** La région R_n associé au test simple φ_n est dite **région de rejet** / critique de φ_n .

Lorsqu'on procède à un test, on décide d'accepter l'hypothèse H_0 si l'évènement $\{\varphi_n = 0\}$ est réalisé ou de la rejeter si $\{\varphi_n = 1\}$ est réalisé. On peut avoir raison de 2 manières différentes :

1. On observe $\{\varphi_n = 0\}$ et on a $F(x_0) = F_0(x_0)$.
2. On observe $\{\varphi_n = 1\}$ et on a $F(x_0) \neq F_0(x_0)$.

On peut également se lopper de 2 manières différentes, en inversant les 2 items précédentes.

On va essayer de minimiser **ensemble** ces deux erreurs. Pour cela, on définit $F(x_0) = F_0(x_0)$ et $F(x_0) \neq F_0(x_0)$ précisément.

Soit $\mathcal{F} = \{F \in \mathcal{F}, F \text{ f}^\circ \text{ de rép.}\}$. Posons $\mathcal{F}_0 = \{F \in \mathcal{F}, F(x_0) = F_0(x_0)\}$. Alors l'hypothèse H_0 se traduit par le sous-ensemble des paramètres de \mathcal{F}_0 , et l'hypothèse alternative H_1 pour $\mathcal{F} / \mathcal{F}_0$.

Définition 2.2.6 Soit $\alpha \in [0, 1]$. Le test φ_n est de niveau α si :

$$\sup_{F \in \mathcal{F}} \mathbb{P}_F(\varphi = 1) \leq \alpha \quad (2.15)$$

Autrement dit, la probabilité de rejeter l'hypothèse H_0 (i.e. d'observer $\{\varphi_n = 1\}$, alors qu'elle est vraie ($F \in \mathcal{F}_0$)) est inférieure à α .

■ **Exemple** Probabilité d'envoyer un innocent en prison. ■

Définition 2.2.7 La puissance du test φ_n est l'application de $\mathcal{F} / \mathcal{F}_0$ dans $[0, 1]$ définie par la probabilité du rejet, soit :

$$F \in \mathcal{F} / \mathcal{F}_0 \rightsquigarrow \mathbb{P}_F(\varphi_n = 1) \quad (2.16)$$

On parle indifféremment de "puissance du test φ_n " ou bien de fonction d'erreur de seconde espèce, définie par :

$$F \in \mathcal{F} / \mathcal{F}_0 \rightsquigarrow 1 - \mathbb{P}_F(\varphi_n = 1) \quad (2.17)$$

A partir d'estimateurs et des intervalles de confiance de \mathbb{P} , de couverture (asy. ou non) $1 - \alpha$, la construction du test est naturelle. On a, $\forall F \in \mathcal{F}$, $\mathbb{P}_F[F(x_0) \in J_{n,\alpha}] \xrightarrow[n \rightarrow \infty]{} 1 - \alpha$. Cela suggère la règle suivante :

Propriété 2.2.12 On accepte H_0 si $F_0(x_0) \in J_{n,\alpha}$, et on la rejette sinon.

Soit $\alpha \in [0, 1]$. Le test $\varphi_n = \varphi_{n,\alpha}$ de l'hypothèse H_0 contre l'alternative H_1 est définie par la zone de rejet $R_{n,\alpha} = \{F_0(x_0) \notin J_{n,\alpha}\}$, et il est asymptotiquement de niveau α (ou erreur de première espèce).

De plus, pour tout point de l'alternative $F \in \mathcal{F} / \mathcal{F}_0$, on a :

$$\mathbb{P}_F(\varphi_{n,\alpha}) = \mathbb{P}_F((x_1, \dots, x_n) \notin R_{n,\alpha}) \xrightarrow[n \rightarrow \infty]{} 0 \quad (2.18)$$

R Autrement dit, l'erreur de première espèce est asymptotiquement plus petite que α et l'erreur de seconde espèce tend vers 0 pour $n \rightarrow \infty$. Ce test est donc **consistant**.

2.3 Estimation uniforme

On a travaillé sur un point local, qu'en est-il en général ? (i.e. $F(x_0) \Rightarrow \forall x \in \mathbb{R}, \hat{F}_n(x)$) ?

2.3.1 Estimation uniforme

Théorème 2.3.1 — Glivenko-Cantelli. Soient (X_1, \dots, X_n) n variables aléatoires réelles i.i.d., de loi F et $\hat{F}_n(x) = \frac{1}{n} \sum \mathbf{1}_{x_i \leq x}$. Alors :

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{p.s.} 0 \quad (2.19)$$

On dit que $\hat{F}_n(x)$ est un estimateur (sup) uniformément (p.s.) consistant.

Théorème 2.3.2 — Kolgomorov-Smirnov. Si F continue, alors :

$$\sqrt{n} \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{L} \mathcal{B} \quad (2.20)$$

Ce \mathcal{B} est une variable aléatoire dont la loi ne dépend **pas** de F !

R [HP] Ici, $\mathcal{B} = \sup_{t \in \mathbb{R}} B(t)$, où $B(t)$ est un processus aléatoire, appelé "Pont Brownien".

Lemme Soit U_1, \dots, U_n n variables aléatoires i.i.d. $\sim U([0, 1])$. On note $\hat{G}_n(x) = \sum \mathbf{1}_{U_i \leq x}$. Si F continue, on a :

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| = \sup_{u \in \mathbb{R}} |\hat{G}_n(u) - u| \quad (2.21)$$

R On peut en réalité restreindre u à $[0, 1]$.

Démonstration. Posons $U_i = F_{x_i}(x_i)$. Alors par la méthode d'inversion totale, on a $U_i \sim U([0, 1])$. On peut donc écrire $\hat{F}_n(x) = \frac{1}{n} \sum \mathbf{1}_{F_i(x_i) \leq F_i(x)} = \frac{1}{n} \sum \mathbf{1}_{U_i \leq \text{pt en loi}} = \hat{G}_n(x)$. ■

2.3.2 Intervalle de confiance uniforme

Propriété 2.3.3 La région $\{J_{n,\alpha}(x), x \in \mathbb{R}\} = \{[\hat{F}_n(x) \pm \frac{q_{1-\alpha}}{\sqrt{n}}], x \in \mathbb{R}\}$ est une région de confiance asymptotique. Ainsi :

$$\mathbb{P}[\forall x \in \mathbb{R}, F(x) \in J_{n,\alpha}(x)] \rightarrow 1 - \alpha \quad (2.22)$$

Démonstration.

$$\mathbb{P}[\forall x \in \mathbb{R}, F(x) \in J_{n,\alpha}(x)] = \mathbb{P}[\sup_{x \in \mathbb{R}} \sqrt{n} |\hat{F}_n(x) - F(x)| \leq q_{1-\alpha}] \rightarrow \mathbb{P}(\mathcal{B} \leq q_{1-\alpha}) = 1 - \alpha$$

■

R Attention, $J_{n,\alpha}(x)$ est différent de $J_{n,\alpha}$! Le premier est une fonction (cf. TCL), tandis que l'autre dépend d'un point fixé.

Conclusion : $\text{TCL} \rightarrow I_{n,\alpha} \rightarrow \text{Test}$.

\mathcal{A} : Asymptotique (n infini), ou non asymptotique (n fixé)

\mathcal{B} : Ponctuelle (x fixé) ou uniforme ($\forall x \in \mathbb{R}$)

2.4 Estimation de fonctionnelles - Méthode de substitution

On a rencontré 2 situations :

1. "Estimation locale" de F en x_0 : on a vu $T_{x_0}(F) = F(x_0)$.
2. "Estimation globale" de F i.e. une fonctionnelle de type $T_x(F) = F(x), x \in \mathbb{R}$.

On peut considérer des fonctionnelles plus générales, les p cumulées :

1. Une fonctionnelle linéaire $T(F) = \int_{\mathbb{R}} g(x) dF(x)$, avec g donné. Par exemple, si $g(x) = x$, alors $T(F) = \int_{\mathbb{R}} x dF(x) = \mu(F)$.

2. Une combinaison de fonctionnelles linéaires :

- Variance : $T(F) = \int_{\mathbb{R}} (x - \mu(F))^2 dF(x) = \sigma^2(F)$
- Coefficient d'asymétrie : $\alpha(F) = \frac{1}{(\sigma^2(F))^{3/2}} \int_{\mathbb{R}} (x - \mu(F))^3 dF(x)$
- Kurtosis : $K(F) = \frac{\int_{\mathbb{R}} (x - \mu(F))^4 dF(x)}{(\sigma^2(F))^2}$

3. Une fonctionnelle non linéaire, comme le quantile de F au niveau $\alpha \in [0, 1]$.

Définition 2.4.1 L'estimateur par substitution (ou play-in) de $T(F)$ est :

$$\hat{T}_n = \hat{T}_n(x_1, \dots, x_n) = T(\hat{F}_n) \quad (2.23)$$

Ici, \hat{F}_n est la f° de répartition empirique associée à F .

Propriété 2.4.1 Si $T(F) = h(\int_{\mathbb{R}} g(x) dF(x))$, où $\int |g(x)| dF(x) < +\infty$ et h une fonction réelle continue, alors $T(\hat{F}_n) \xrightarrow{p.s.} T(F)$.

Démonstration. Si $T(F) = h(\int_{\mathbb{R}} g(x) dF(x))$, alors $T(\hat{F}_n) = h(\frac{1}{n} \sum g(X_i))$, et on a

$$\frac{1}{n} \sum g(X_i) \xrightarrow{p.s.} \frac{1}{n} \mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x) dF(x) \quad (2.24)$$

Cette convergence est presque sûre par la loi forte des grands nombres. Ainsi, grâce au continuous mapping theorem (puisque h continue !), on a :

$$T(\hat{F}_n) = h(\frac{1}{n} \sum g(X_i)) \xrightarrow{p.s.} h(\int_{\mathbb{R}} g(x) dF(x)) = T(F) \quad (2.25)$$

■

Pour pouvoir travailler sur des lois et des vitesses de convergence, il faut le TCL !

Théorème 2.4.2 — TCL. Soit $T(F) = h(\int_{\mathbb{R}} g(x) dF(x))$. Si $h \in C^1(\mathbb{R})$ et $\mathbb{E}(g(x)^2) = \int g(x)^2 dF(x) < +\infty$, alors :

$$\sqrt{n}(T(\hat{F}_n) - T(F)) \rightarrow N(0, v(F)), v(F) = h'(\mathbb{E}(g(X)))^2 V(g(X)) \quad (2.26)$$

Démonstration.

$$\begin{aligned} \sqrt{n}(\int_{\mathbb{R}} g(x) d\hat{F}_n(x) - \int_{\mathbb{R}} g(x) dF(x)) &= \sqrt{n}(\frac{1}{n} \sum g(x_i) - \mathbb{E}[g(X)]) \\ &\rightarrow N(0, V(g(X))) \end{aligned}$$

Par la Δ -méthode,

$$\sqrt{n}(h(\frac{1}{n} \sum g(X_i)) - h(\int_{\mathbb{R}} g(x) dF(x))) \xrightarrow{L} N(0, V(g(X)) h'(\mathbb{E}[g(X)])^2)$$

Ce qui donne $\sqrt{n}(T(\hat{F}_n) - T(F)) \rightarrow N(0, h'(\mathbb{E}(g(X)))^2 V(g(X)))$. ■

Peut-on généraliser en dimension K ? On va voir ça :

Propriété 2.4.3 Soit $h : \mathbb{R}^* \rightarrow \mathbb{R}$ différentiable par produit : $J_h(x) = \nabla h(x) = (\partial_1 h(x), \dots, \partial_n h(x)), x \in \mathbb{R}^*$. Alors :

$$T(F) = h(\int_{\mathbb{R}} g_1(x) dF(x), \dots, \int_{\mathbb{R}} g_n(x) dF(x)) \quad (2.27)$$

Corollaire 2.4.4 Si $T(F)$ vérifie (2.4.3), que $\int g_i(x)^2 dF(x) < +\infty$, et que $\forall i \in [1, K]$, h est différentiable, alors :

$$\sqrt{n}(T(\hat{F}_n) - T(F)) \xrightarrow{L} N(0, v(F)) \quad (2.28)$$

Avec $v(F) = J_h(g)\Sigma_g J_h(g)$ et $g = (\mathbb{E}[g_1(x)], \dots, \mathbb{E}[g_n(x)])$.

Ainsi,

$$\forall i, j \in [1, K] \Sigma_{g_{ij}} = \mathbb{E}[(g_i(X) - \mathbb{E}[g_i(X)])(g_j(X) - \mathbb{E}[g_j(X)])] \quad (2.29)$$

3. Méthode d'estimation en densité

3.1 Introduction

3.1.1 Notations et hypothèses

On donne un échantillon (X_1, \dots, X_n) . Les X_i sont des variables aléatoires **i.i.d.** et on suppose que leur loi sont dans une famille paramétrique.

3.1.2 Familles paramétriques classiques

- **Exemple** — $\mathbb{R}_\theta^2 = \text{Exp}(\lambda), \lambda \in \mathbb{R}^* = \Theta(d=1)$
- $\mathbb{R}_\theta^2 = N(\mu, \sigma^2), \mu, \sigma \in \mathbb{R} * \mathbb{R}^*$

■

Dans ce contexte, on cherche à construire des estimateurs $\hat{\theta}_n$ de θ , variant avec n , tel que $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$.

- R** Pour nos estimateurs paramétriques, nous allons considérer μ la mesure de Lebesgue et u la mesure de comptage discrète.

On prend une fonction test φ . Ainsi,

$$\begin{aligned}\mathbb{E}[\varphi(\hat{\theta}_n)] &= \mathbb{E}_\theta[\varphi(\hat{\theta}_1, \dots, \hat{\theta}_n)] = \int_{\mathbb{R}^n} \varphi(\hat{\theta}_n(x_1, \dots, x_n)) \\ &= \mathbb{P}_\theta(dx_1) \cdots \mathbb{P}_\theta(dx_n) \\ &= \int_{\mathbb{R}^n} \varphi(\hat{\theta}_n(x_1, \dots, x_n)) \prod_{i=1}^n f(\theta, x_i) \mu(dx_1) \cdots \mu(dx_n)\end{aligned}$$

Si μ est une mesure de Lebesgue en a :

1. $\mathbb{E}_\theta(\varphi(\hat{\theta}_n(x_1, \dots, x_n))) \prod f(\theta, x_i) dx_1, \dots, dx_n$

Si μ est la mesure discrète de comptage sur $A \subset \mathbb{R}$:

2. $\mathbb{E}_\theta(\varphi(\hat{\theta}_n)) = \sum_{x_1, \dots, x_n} \varphi(\hat{\theta}_n(x_1, \dots, x_n)) \prod f(\theta, x_i).$

3.2 Méthode des moments

3.2.1 En dimension 1

On suppose $\theta \in \mathbb{R}$ et $g : \mathbb{R} \rightarrow \mathbb{R}$ tel que $\theta \rightsquigarrow n(\theta) = \mathbb{E}_\theta[g(X)]$ existe, soit croissante et C^0 . Alors :

$$\theta = n^{-1}(\mathbb{E}[g(X)]), \theta \in \Theta \quad (3.1)$$

En remplaçant la moyenne théorique par celle empirique dans l'équation 3.2.1, on obtient :

Propriété 3.2.1 — Estimateur des moments.

$$\hat{\theta}_n = n^{-1} \left(\frac{1}{n} \sum g(X_i) \right) \quad (3.2)$$

R Le moment vient de $g(x) = x^k, k \geq 1$.

Proposition 3.2.2 Si on a $\mathbb{E}[|g(X)|] < +\infty$, on a $\hat{\theta}_n \xrightarrow{p.s.} \theta$. De plus, si $\forall \theta \in \Theta, \mathbb{E}[g(X)^2] < +\infty$, avec n dérivable, alors par la Δ -méthode, on a :

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{L} N(0, \frac{1}{n'(\theta)^2} V(g(X)))$$

3.2.2 En dimension d

On prend cette fois-ci $\Theta \subset \mathbb{R}^d, d \geq 1$. Soit $g_l : \mathbb{R} \rightarrow \mathbb{R}, l \in [1, d]$ tel que $x \rightsquigarrow (g_1(x), \dots, g_d(x))$, de sorte à avoir le système d'équation :

$$m_p(x) = \mathbb{E}[g_l(x)] = \int_{\mathbb{R}} g_l(x) dF_\theta(x), l \in [1, d]$$

Définition 3.2.1 On note $m(\theta) = \mathbb{E}[g(X)] = (\mathbb{E}[g_1(X)], \dots, \underbrace{\mathbb{E}[g_d(X)]}_{\int g_d(x) dF(x)})$. On utilise la représentation $\theta = m^{-1}(m_1(\theta), \dots, m_d(\theta))$ pour estimer $\hat{\theta}_n = m^{-1}(\frac{1}{n} \sum g_1(x_i), \dots, \frac{1}{n} \sum g_d(x_i))$.

Proposition 3.2.3 Si m continue, inversible, alors $\hat{\theta}_n \xrightarrow{p.s.} \theta$. De plus, si m^{-1} différentiable et si $\mathbb{E}_\theta(g_l(X)^2) < +\infty$, on a la convergence en loi suivante :

$$\sqrt{n}(\hat{\theta}_n - \theta) \rightarrow N(0, V(\theta)) = N(0, J_{m^{-1}} \Sigma_m(\theta)^t J_{m^{-1}}) \quad (3.3)$$

Démonstration.

$$\left(\frac{1}{n} \sum g_1(X_i), \dots, \frac{1}{n} \sum g_d(X_i) \right) \xrightarrow{p.s.} (\mathbb{E}[g_1(X)], \dots, \mathbb{E}[g_d(X)]) = m(\theta)$$

Par continuité de m^{-1} , on conserve la convergence i.e.

$$\hat{\theta}_n = m^{-1} \left(\frac{1}{n} \sum g_1(X_i), \dots, \frac{1}{n} \sum g_d(X_i) \right) \xrightarrow{p.s.} m^{-1}(\mathbb{E}[g_1(X)], \dots, \mathbb{E}[g_d(X)]) = m^{-1}(m(\theta)) = \theta$$

En utilisant la Δ -méthode multidimensionnelle, on a :

$$\sqrt{n} \left(\left(\frac{1}{n} \sum g_1(x_i), \dots, \frac{1}{n} \sum g_d(x_i) \right) - m(\theta) \right) \xrightarrow{L} N(0, \Sigma_m(\theta))$$

Enfin, on applique la Δ -méthode en dimension d avec $J_{m^{-1}}$. ■

3.3 Moments généralisés, M-estimateur et Z-estimateur

3.3.1 Z-estimateur

En dimension 1, prenons $\theta \in \mathbb{R}$ et le théorème 3.2.1 (sur des bonnes propriétés de la fonction m !). Cette dernière est basée sur l'écriture $m(\theta) = \int_{\mathbb{R}} g(x) \mathbb{P}_{\theta}(dx)$ pour une certaine fonction g .

Autrement dit, $\forall \theta \in \Theta$, et à g fixé,

$$\int_{\mathbb{R}} (m(\theta) - g(x)) \mathbb{P}_{\theta}(dx) = 0 \quad (3.4)$$

Théorème 3.3.1 En dimension d , soit $\Theta \subset \mathbb{R}^d$, $\Phi : \Theta \times \mathbb{R} \rightarrow \mathbb{R}$ tel que $\forall \theta \in \Theta$,

$$\int_{\mathbb{R}} \Phi(\theta, x) \mathbb{P}_{\theta}(dx) = 0 \quad (3.5)$$

Ainsi, ce théorème est un cas général du théorème 3.3.1, en posant $\Phi(\theta, x) = m(\theta) - g(x)$.

R Un estimateur possible est $\frac{1}{n} \sum \phi(\hat{\theta}_n, X_i) = 0$.

Définition 3.3.1 Étant donné $\Phi : \Theta \times \mathbb{R} \rightarrow \mathbb{R}$ satisfaisant 3.3.1. On appelle Z-estimateur (ou estimateur GMM) tout estimateur $\hat{\theta}_n$ satisfaisant 3.3.1.

Soit $\Phi : \Theta \times \mathbb{R} \rightarrow \mathbb{R}^d$ tel que $\Phi \mapsto (\Phi_1, \dots, \Phi_d)$. Le paramètre θ inconnu est la solution d'un système d'équations. On a, pour $l \in [1, d]$,

$$\int_{\mathbb{R}} \phi_l(\theta, x) \mathbb{P}_{\theta}(dx) = 0 \quad (3.6)$$

On peut en construire un Z-estimateur (x_1, \dots, x_n) tel que :

$$\frac{1}{n} \sum \Phi_l(\hat{\theta}_n, X_i) = 0, l \in [1, d] \quad (3.7)$$

3.3.2 M-estimateur

Soit $\psi : \Theta \times \mathbb{R} \rightarrow \mathbb{R}$ une application tel que $\forall \theta \in \Theta \subset \mathbb{R}^d$, $d \geq 1$, la fonction

$$a \rightsquigarrow \mathbb{E}_{\theta}[\psi(a, X)] = \int_{\mathbb{R}} \psi(a, x) \mathbb{P}_{\theta}(dx) \quad (3.8)$$

admette un maximum en $a = \theta$.

Ainsi, chercher un estimateur de θ consiste à construire une version empirique de 3.3.2.

Définition 3.3.2 On appelle M-estimateur associé au contraste ψ tout estimateur $\hat{\theta}_n$ qui satisfait l'équation

$$\frac{1}{n} \sum \psi(\hat{\theta}_n, X_i) = \max_{a \in \Theta} \frac{1}{n} \sum \psi(a, X_i) \quad (3.9)$$

R Si le paramètre θ est de dimension $d = 1$, alors en prenant $\phi(a, x) = \frac{\partial_1}{\partial x} \psi(a, x)$, on a :

$$\sum_{i=1}^n \frac{\partial_1}{\partial a} \psi(\hat{\theta}_n, X_i) = \sum \phi(\hat{\theta}_n, X_i) = 0 \quad (3.10)$$

On peut ainsi voir le M-estimateur comme un Z-estimateur.

3.3.3 Convergence de M-estimateur et Z-estimateur

Pour une fonction contraste $\psi : \Theta \times \mathbb{R} \rightarrow \mathbb{R}$, on définit $M_n(a) = \frac{1}{n} \sum \psi(a, X_i)$, $a \in \Theta$ et pour $\theta \in \Theta$, $M(a, \theta) = \mathbb{E}_\theta[\psi(a, X)]$.

Proposition 3.3.2 — Admis. "Sous des conditions de régularité pour $M(a, \theta)$ ", alors le M -estimateur $\hat{\theta}_n \xrightarrow{p.s.} \theta$. Analytiquement, "sous des conditions de régularité pour $Z(a, \theta) = \mathbb{E}_\theta[\phi(a, X)]$ ", le Z -estimateur $\hat{\theta}_n \xrightarrow{p.s.} \theta$.

3.3.4 Loi limite des Z-estimateurs et M-estimateurs

On se place sous les hypothèses suivantes :

1. Pour tout $\theta \in \Theta$, il existe un voisinage $V(x)$ tel ue $\forall a \in V(x)$, on a $|\partial_x^2 \phi(a, x)| < g(x)$, où $\mathbb{E}[g(X)] < +\infty$.
2. $\forall \theta \in \Theta$, $\mathbb{E}_\theta[\phi(\theta, X)] = 0$, $\mathbb{E}_\theta[\phi(\theta, x)^2] < +\infty$ et $\mathbb{E}_\theta[\partial_\theta \phi(\theta, x)] \neq 0$.

Théorème 3.3.3 Si ϕ vérifie les hypothèses 1 et 2, alors $\hat{\theta}_n$ est un Z-estimateur tel que :

$$\sqrt{n}(\hat{\theta}_n - \theta) \rightarrow N(0, V_\phi(\theta)), V_\phi(\theta) = \frac{\mathbb{E}[\phi(\theta, x)^2]}{(\mathbb{E}[\partial_\theta \phi(\theta, x)])^2} \quad (3.11)$$

Démonstration. Notons $Z_n(a) = \frac{1}{n} \sum \phi(a, X_i)$ pour $a \in \Theta$.

Introduisons les notations $Z'_n(a) = \partial_a Z_n(a)$ et $Z''_n(a) = \partial_a^2 Z_n(a)$. On effectue un développement de Taylor à l'ordre 2 de $Z_n(a)$ au $V(\theta)$.

$$Z_n(\hat{\theta}_n) = 0 = Z_n(\theta) + (\hat{\theta}_n - \theta)Z'_n(\hat{\theta}_n) + \frac{1}{2}(\hat{\theta}_n - \theta)^2 Z''_n(\tilde{\theta}_n) \quad (3.12)$$

Ici, $\tilde{\theta}$ est un point aléatoire entre $\hat{\theta}_n$ et θ . On peut donc réécrire l'équation 3.3.4 :

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{-\sqrt{n}Z_n(\theta)}{Z'_n(\hat{\theta}_n) + \frac{1}{2}(\hat{\theta}_n - \theta)Z''_n(\tilde{\theta}_n)} \quad (3.13)$$

On sait que sous $\mathbb{P}_\theta(dx)$, $\phi(\theta, X_i)$ sont i.i.d. à variance finie (hyp. 2). On peut ainsi appliquer le TCL sur $\phi(\theta, X_i)$:

$$\underbrace{-\sqrt{n}(Z_n(\theta))X}_{\frac{1}{n} \sum \phi(\theta, X_i)} \xrightarrow{L} N(0, \mathbb{E}_\theta(\phi(\theta, X)^2)) \quad (3.14)$$

μ est nul d'après l'hypothèse 1. Considérons maintenant le dénominateur $Z'_n(\theta) = \frac{1}{n} \sum \partial_\theta \phi(\theta, X_i) \xrightarrow{p.s.} \mathbb{E}[\partial_\theta \phi(\theta, X_i)] \neq 0$. La difficulté est de démontrer que $\frac{1}{2}(\hat{\theta}_n - \theta)Z''_n(\tilde{\theta}_n) \xrightarrow{p.s.} 0$.

Ainsi, le dénominateur converge p.s. :

- $D_n \xrightarrow{p.s.} \mathbb{E}[\partial_\theta \phi(\theta, X)] \neq 0$
- $T_n \rightarrow N(0, \sigma^2)$
- $D_n \xrightarrow{p.s.} c$

Ainsi, d'après le théorème de Slutsky, on peut combiner les convergences en L et $p.s$. Finalement, on a réussi à remplacer ϕ par l'expression d'une dérivée qui s'annule. Ainsi, cela revient à dire que le $\hat{\theta}_n$ est le maximum mais également le point où $\partial_a \psi = 0$. ■

3.4 Maximum de vraisemblance

3.4.1 Principe du maximum de vraisemblance

Soit l'expérience $\varepsilon = (X_1, \dots, X_n)$, de mesure μ sur \mathbb{R} , et on note $\{f(\theta, \cdot), \theta \in \Theta\}$ la famille de densité par rapport à μ .

Pour toute fonction test ϕ ,

$$\int \phi(x) \mathbb{P}_\theta(dx) = \int \phi(x) f(\theta, x) \mu(dx) = \int \phi(x) f(\theta, \mu) dx \quad (3.15)$$

Définition 3.4.1 On appelle fonction de vraisemblance associé à l'expérience ε l'application :

$$\theta \in \Theta \rightsquigarrow L_n(\theta, X_1, \dots, X_n) = \prod_{i=1}^n f(\theta, X_i) \quad (3.16)$$

R La fonction L_n est une fonction aléatoire mesurable, observable sur l'expérience ε .

Définition 3.4.2 — Estimateur de vraisemblance. On appelle estimateur de maximum de vraisemblance toute valeur $\hat{\theta}_n^m$ tel que :

$$\hat{\theta}_n^m \in \arg \max_{\theta \in \Theta} L_n(\theta, X_1, \dots, X_n) \quad (3.17)$$

Autrement dit :

$$L_n(\hat{\theta}_n^{mv}, X_1, \dots, X_n) = \max_{\theta \in \Theta} L_n(\theta, X_1, \dots, X_n) \quad (3.18)$$

R La maximisation se fait sur le paramètre θ (et donc $\hat{\theta}_n^{mv}$) et non les X_i ! De plus, cet estimateur de vraisemblance peut ne pas exister, et ne pas être unique.

On va prendre la définition précédente (3.4.2) et passer au log népérien par praticité :

Définition 3.4.3 — Log-vraisemblance. L'application

$$\theta \in \Theta \rightsquigarrow \ln(\theta, X_1, \dots, X_n) = \frac{1}{n} \ln(L_n(\theta, X_1, \dots, X_n)) = \frac{1}{n} \sum \ln(f(\theta, X_i)) \quad (3.19)$$

est bien définie si $f(\theta, \cdot) > 0$, et elle est appelée fonction de log-vraisemblance.

En fixant $\ln(0) = 0$, on conserve les définitions précédentes.

3.4.2 Principe de vraisemblance pour 2 points

Considérons une famille de lois à deux points $\Theta = (\theta_1, \theta_2) \subset \mathbb{R}$, où \mathbb{P}_{θ_1} et \mathbb{P}_{θ_2} sont deux lois discrètes, μ mesure de comptage et la fonction de densité du modèle est donné par :

$$f(\theta, x) = \mathbb{P}_\theta[X = x], x \in M, \theta \in [\theta_1, \theta_2] \quad (3.20)$$

A priori (avant ε), si les observations suivent la loi \mathbb{P}_θ (avec $\theta = \theta_1$ ou θ_2), alors la probabilité d'observer $\{X_1 = x_1, \dots, X_n = x_n\}$ est donné par :

$$\mathbb{P}_\theta\{X_1 = x_1, \dots, X_n = x_n\} = \prod_{i=1}^n \mathbb{P}_\theta[X_i = x_i] = \prod_{i=1}^n f(\theta, x_i) \quad (3.21)$$


A posteriori, on dispose maintenant des observations (X_1, \dots, X_n) . Supposons que l'on observe $\{\prod f(\theta_1, X_i) > \prod f(\theta_2, X_i)\}$, i.e.

$$L_n(\theta_1, X_1, \dots, X_n) > L_n(\theta_2, X_1, \dots, X_n) \quad (3.22)$$

D'après l'équation 3.4.2, nous pouvons interpoler l'équation 3.4.2 de la façon suivante :

A posteriori, la probabilité d'avoir observé (X_1, \dots, X_n) est plus grande sous \mathbb{P}_{θ_1} que sous \mathbb{P}_{θ_2} . Cela nous suggère de "suspecter" que la loi des observations suit \mathbb{P}_{θ_1} plutôt que \mathbb{P}_{θ_2} . On dira donc que la valeur " θ_1 est plus vraisemblable que la valeur θ_2 ". On a donc :

$$\hat{\theta}_n^{mv} = \theta_1 \mathbf{1}_{\{L_n(\theta_1, X_1, \dots, X_n) > L_n(\theta_2, X_1, \dots, X_n)\}} + \theta_2 \mathbf{1}_{\{L_n(\theta_1, X_1, \dots, X_n) < L_n(\theta_2, X_1, \dots, X_n)\}} \quad (3.23)$$

 Attention, il n'y a pas d'unicité si $L_n(\theta_1) \equiv L_n(\theta_2)$.

3.4.3 Passage à une famille de lois quelconque

De manière générale, si $\Theta \subset \mathbb{R}^d$, $d \geq 1$ est un ensemble arbitraire, alors la valeur

$$\hat{\theta}_n^{mv} \in \arg \max_{\theta \in \Theta} L_n(\theta, X_1, \dots, X_n) \quad (3.24)$$

est la plus vraisemblable (si elle est bien définie !).

$\{\mathbb{P}_\theta, \theta \in \Theta\}$ est absolument continue par rapport à la mesure de Lebesgue μ , on peut ainsi "reproduire" le raisonnement précédent :

$$\mathbb{P}_\theta[X_1 = x_1, \dots, X_n = x_n] = \prod \mathbb{P}_\theta[X_i = x_i] = \prod f(\theta, x_i) \quad (3.25)$$

En prenant $v(x)$ un "petit" voisinage de x , on a :

$$\mathbb{P}_\theta[X_1 \in v(x_1), \dots, X_n \in v(x_n)] = \prod \mathbb{P}_\theta[X_i \in v(x_i)] \quad (3.26)$$

Alors :

$$\mathbb{P}_\theta[x \in v(x)] = \int_{v(x)} f(\theta, u) du \sim f(\theta, x) |v(x)| \quad (3.27)$$

dans la limite où $|v(x)| \rightarrow 0$, $|v(x)|$ désignant la mesure de Lebesgue de $v(x)$. Ainsi la probabilité de l'évènement $\{X_1 \in v(x_1), \dots, X_n \in v(x_n)\}$ est "proportionnelle" à $\prod f(\theta, x_i)$, et ceci indépendamment de θ (si l'on accepte l'approximation heuristique en 3.4.3).


3.4.4 Maximum de vraisemblance et M-estimateur

Soit l'application $\theta \mapsto L_n(\theta)$ différentiable, alors une CN que doit satisfaire l'estimateur de maximum de vraisemblance $\hat{\theta}_n^{mv}$ est l'annulation du gradient i.e. $\hat{\theta}_n^{mv}$ est un point critique en L_n :

$$\nabla_\theta L_n(\theta, X_1, \dots, X_n) = 0 \quad (3.28)$$

Ainsi cela nous fournit un système d'équations si $\Theta \subseteq \mathbb{R}^d$:

$$\nabla_\theta \ln(\theta, X_1, \dots, X_n) \Big|_{\theta = \hat{\theta}_n^{mv}} = 0 \quad (3.29)$$

 Si $d = 1$, on appelle cela une équation de vraisemblance, et si $d > 1$, il s'agit d'un système de vraisemblance.

Définition 3.4.4 On appelle racine de l'équation de vraisemblance tout estimateur $\hat{\theta}_n^{mv}$ solution de l'équation / du système 3.29.

Théorème 3.4.1 Le maximum de vraisemblance **est** un M -estimateur (p.s., TCL).

Retournons dans le cadre de la M -estimation. Posons $\psi(a, x) = \ln f(a, x)$, f densité, $(a, x) \in \Theta \times \mathbb{R}$.

Alors, si $\hat{\theta}_n^{mv}$ existe, et satisfait $\hat{\theta}_n^{mv} \in \arg \max \sum \psi(a, X_i)$, on peut l'interpréter comme un M -estimateur de la fonction $\psi(a, x)$. En effet, $a = \theta$ maximise :

$$a \rightsquigarrow \int_{\mathbb{R}} \psi(a, x) \mathbb{P}_{\theta}(dx) = \int_{\mathbb{R}} \ln(f(a, x)) f(\theta, x) dx \quad (3.30)$$

Si, pour tout point $\theta \in \Theta$, la fonction $\theta \rightsquigarrow \ln(f(\theta, x))$ est différentiable presque partout, alors on a aussi une interprétation de $\hat{\theta}_n^{mv}$ comme un Z -estimateur associé à la fonction $\phi(\theta, x) = \partial_{\theta} \ln(f(\theta, x)) = \frac{\partial_{\theta} f(\theta, x)}{f(\theta, x)}$, $\theta \in \Theta, x \in \mathbb{R}$.



Une généralisation pour $d \geq 1$ est possible !

4. Méthode d'estimation en régression

4.1 Régression linéaire simple

Définition 4.1.1 On appelle modèle de régression linéaire simple l'expérience statistique engendrée par les v.a. Y_i à valeurs dans \mathbb{R} et par le "design" où $x_i \in \mathbb{R}$ tel que $Y_i = \theta_0 + \theta_1 x_i + \xi_i$, $i \in \llbracket 1, n \rrbracket$.

- Les paramètres inconnus sont $\theta = {}^t(\theta_0, \theta_1) \in \Theta = \mathbb{R}^2$.
- Les ξ sont des "bruits", ayant une espérance nulle et une homoscedasticité (σ^2).

4.1.1 Droite de régression

On peut parler de modèle de RLS à variance connue / inconnue. Les paramètres θ_0 représentent l'ordonnée à l'origine et θ_1 le coefficient directeur. Ainsi,

$$y = r(\theta, x) = \theta_0 + \theta_1 x$$

Si $\hat{\theta}_n$ est un estimateur de $\theta = (\theta_1, \theta_2)$, on note $x \rightsquigarrow r(\hat{\theta}_n, x) = \hat{\theta}_0 + \hat{\theta}_1 x$ l'estimateur de la fonction de régression. On note également $\hat{Y}_i = r(\hat{\theta}_n, x_i)$ la valeur prédite par la RLS, mais aussi $\hat{\xi}_i = Y_i - \hat{Y}_i$ le bruit / résidu.

Définition 4.1.2 — Residual Square Source.

$$\|\hat{\xi}\|^2 = \sum_{i=1}^n \hat{\xi}_i^2 = \sum (\hat{Y}_i - Y_i)^2 \quad (4.1)$$

R La RSS mesure l'erreur (en norme $\|\cdot\|^2$) entre les observations et les prédicteurs.

4.1.2 Moindres carrés et maximum de vraisemblance

Définition 4.1.3 — Estimateur des MC. L'estimateur des moindres carrés $\hat{\theta}_n^{mc}$ est l'estimateur minimisant le RSS.

$$\|\hat{\xi}\|^2 = \sum (Y_i - r(\hat{\theta}_n^{mc}, x_i))^2 = \min_{\theta \in \Theta = \mathbb{R}^2} \sum (Y_i - r(\theta, x_i))^2$$

L'infimum est pris sur toutes les valeurs possibles (i.e. \mathbb{R}^2).

R On considère dans la suite la norme $l(u) = u^2$ et non la norme $l(u) = |u|$.

On va maintenant explicier les points $\hat{\theta}_n^{mc} = (\hat{\theta}_0^{mc}, \hat{\theta}_1^{mc})$.

Propriété 4.1.1 En posant $\bar{x}_n = \frac{1}{n} \sum x_i$ et $\bar{Y}_n = \frac{1}{n} \sum Y_i$, on a :

$$\hat{\theta}_{0,n} = \bar{Y}_n - \hat{\theta}_{1,n}^{mc} \bar{x}_n = \frac{\sum (x_i - \bar{x}_n) Y_i}{\sum (x_i - \bar{x}_n)^2} \quad (4.2)$$

Démonstration. On a $(\theta_0, \theta_1) \rightsquigarrow L_n(\theta_0, \theta_1) = \sum (Y_i - \theta_0 - \theta_1 x_i)^2$. Ainsi ∇ vaut :

$$— \partial_{\theta_0} L_n(\theta_0, \theta_1) = -2 \sum (Y_i - \theta_0 - \theta_1 x_i)^2$$

$$— \partial_{\theta_1} L_n(\theta_0, \theta_1) = -2 \sum (Y_i - \theta_0 - \theta_1 x_i) x_i$$

En annulant le gradient et par linéarité, on obtient le système :

$$\begin{cases} -\sum (Y_i - \theta_0 - \theta_1 x_i) = 0 \\ -\sum Y_i x_i + \theta_0 \sum x_i + \theta_1 \sum x_i^2 = 0 \end{cases}$$

On prend la première équation, en isolant θ_0 :

$$\hat{\theta}_{0,n}^{mc} = \frac{1}{n} \sum Y_i - \frac{\theta_1}{n} \sum x_i = \bar{Y}_n - \theta_1 \bar{x}_n$$

Par plug-in dans la deuxième équation, $\hat{\theta}_{0,n}^{mc} = \bar{Y}_n - \theta_1 \bar{x}_n$, on a $\hat{\theta}_n^{mc}$. ■

Propriété 4.1.2 — RLS. On a $\hat{\theta}_n^{mc}$ vérifiant :

$$\mathbb{E}[\hat{\theta}_n^{mc}] = [\mathbb{E}[\hat{\theta}_{0,n}^{mc}], \mathbb{E}[\hat{\theta}_{1,n}^{mc}]] = (\theta_0, \theta_1) \quad (4.3)$$

Cet estimateur est dit **sans biais** (i.e. $\mathbb{E}[\hat{\theta}_{0,n}^{mc} - \theta_0] = 0$ et $\mathbb{E}[\hat{\theta}_{1,n}^{mc} - \theta_1] = 0$)

De plus, on a une écriture fermée pour $\Sigma_{\hat{\theta}_n^{mc}}$, i.e.

$$\Sigma_{\hat{\theta}_n^{mc}} = \mathbb{E}[(\hat{\theta}_n^{mc} - \theta)(\hat{\theta}_n^{mc} - \theta)^t] = \frac{\sigma^2}{n S_n^2} \begin{pmatrix} \frac{1}{n} \sum x_i^2 & -\bar{x}_n \\ -\bar{x}_n & 1 \end{pmatrix}, s_n^2 = \frac{1}{n} \sum (x_i - \bar{x}_n)^2 \quad (4.4)$$

R La proposition 4.1.2 nous dit que $\mathbb{E}[\hat{\theta}_{0,n}] = \theta_0$ et $\mathbb{E}[\hat{\theta}_{1,n}^{mc} = \theta_1]$, ainsi :

$$\begin{aligned} — V[\hat{\theta}_{0,n}^{mc}] &= \frac{\sigma^2 \frac{1}{n} \sum x_i^2}{n * \frac{1}{n} \sum (x_i - \bar{x}_n)^2} = \frac{\sigma^2}{n} \cdot \frac{\sum x_i^2}{\sum (x_i - \bar{x}_n)^2} \\ — V(\hat{\theta}_{0,n}^{mc}) &= \frac{\sigma^2}{n S_n^2} = \frac{\sigma^2}{n \frac{1}{n} \sum (x_i - \bar{x}_n)^2} = \frac{\sigma^2}{\sum (x_i - \bar{x}_n)^2} \end{aligned}$$

$$— \text{Cov}(\hat{\theta}_0^{mc}, \hat{\theta}_1^{mc}) = \frac{V(\xi_i)\bar{x}_n}{\sum (x_i - \bar{x}_n)^2}$$

Concrètement, augmenter la pente impose une baisse de l'ordonnée à l'origine (et inversement !) i.e. comme une balançoire, si je m'assois d'un côté, je fais remonter l'autre côté.

On ajoute l'hypothèse de normalité sur les résidus au modèle (i.e. les bruits sont i.i.d. et suivent $N(0, \sigma^2)$). On peut donc en faire un modèle à densité pour estimer localement par log-vraisemblance les vecteurs tri-dimensionnels $V(\theta) = (\theta_0, \theta_1, \sigma^2)$.

Propriété 4.1.3 Sous H_N , on a :

$$\hat{\theta}_n^{mv} = (\hat{\theta}_{0,n}^{mv}, \hat{\theta}_{1,n}^{mv}, \hat{\sigma}_n^2) \quad (4.5)$$

Ainsi, on a :

$$\begin{aligned} — & (\hat{\theta}_{0,n}^{mv}, \hat{\theta}_{1,n}^{mv}) = (\hat{\theta}_{0,n}^{mc}, \hat{\theta}_{1,n}^{mc}) \\ — & \hat{\sigma}_n^2 = \frac{1}{n} \sum (\hat{\xi}_i)^2, \hat{\xi}_i = Y_i - \underbrace{r(\hat{\theta}_n^{mc}, x_i)}_{\hat{Y}_i} \end{aligned}$$

Démonstration. $g_\sigma(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{x^2}{2\sigma^2})$ correspond à la densité de la loi normale $N(0, \sigma^2)$ donc on peut calculer la vraisemblance de cette expérience statistique :

$$L_n(\theta_0, \theta_1, \sigma^2, Y_1, \dots, Y_n) = \prod g(Y_i - r(\theta, x_i))$$

On passe au log, ainsi :

$$\ln(\theta_0, \theta_1, \sigma^2, Y_1, \dots, Y_n) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum (Y_i - r(\theta, x_i))^2$$

On peut dériver le log et forcer sa nullité, ce qui donne :

$$\partial_0 = \ln(\dots) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum \underbrace{(Y_i - r(\theta, x_i))^2}_{\|Y - \theta X\|^2} \stackrel{!}{=} 0$$

Ce terme va s'annuler pour $\sigma^2 = \frac{1}{n} \sum \hat{\xi}_i^2 = \frac{1}{n} \sum (Y_i - r(\theta, x_i))^2$.

De même pour $\partial_{\theta_0} l_n$ et $\partial_{\theta_1} l_n$, on retrouve exactement les résultats escomptés. ■

4.1.3 Coefficient de détermination (R^2)

Définition 4.1.4 On pose :

$$\begin{aligned} — & \text{SCT} = \sum (Y_i - \bar{Y}_n)^2 \\ — & \text{SCE} = \sum (\hat{Y}_i - \bar{Y}_n)^2 \\ — & \text{SCR} = \sum (Y_i - \hat{Y}_i)^2 = \sum \hat{\xi}_i^2 \end{aligned}$$

Définition 4.1.5

$$R^2 = \frac{\text{SCE}}{\text{SCT}} = \frac{\|\hat{Y} - \bar{Y}_n \mathbf{1}\|}{\|Y - \bar{Y}_n \mathbf{1}\|} \quad (4.6)$$

R On peut analyser le R^2 comme le \cos^2 entre Y et son projeté orthogonal.

4.2 Régression linéaire multiple

On généralise le modèle de RLS en autorisant les designs X à être vectoriels.

Définition 4.2.1 On considère l'expérience statistique $((x_1, Y_1), \dots, (x_n, Y_n))$, avec :

$$Y_i = {}^t\theta x_i + \xi_i, i = 1, \dots, n \quad (4.7)$$

Y_i sont les variables à valeurs dans \mathbb{R} , x_i sont les variables explicatives à valeurs dans \mathbb{R}^k et ξ le bruit. Ainsi $\theta \in \Theta \subset \mathbb{R}^k$ (vecteur de paramètres à expliquer).

En notant $M = \begin{pmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & & \vdots \\ x_{n,1} & \cdots & x_{n,n} \end{pmatrix}$, on peut vectorialiser le problème, en écrivant $Y = M\theta + \xi$,
où $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ et $\xi = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix}$. Ce bruit blanc a une espérance nulle, et $\mathbb{E}[\xi^t \xi] = \sigma^2 I_n$.

R Attention, la covariance de (X, X^2) ne veut rien dire ici (car X et X^2 sont orthogonaux).

4.2.1 Moindres carrés

L'estimateur $\hat{\theta}_n^{mc}$ minimise la somme des carrés des résidus (SCR) :

$$\sum_{i=1}^n (Y_i - {}^t(\hat{\theta}_n^{mc}) x_i) = \min_{\theta \in \mathbb{R}^n} \sum_{i=1}^n (Y_i - {}^t\theta x_i) \quad (4.8)$$

Définition 4.2.2 L'estimateur $\hat{\theta}_n^{mc}$ correspond à $\hat{\theta}_n^{mc} \in \arg \min_{\theta \in \mathbb{R}^n} \sum_{i=1}^n (Y_i - {}^t\theta x_i)^2$.

Propriété 4.2.1 — MC por la RLM. Soit $M \in M(\mathbb{R})$ tel que tMM soit **inversible**. Alors il existe un **unique** estimateur des MC, s'écrivant :

$$\hat{\theta}^{mc} = ({}^tMM)^{-1} {}^tMY \quad (4.9)$$

Démonstration. Soit $\theta \rightsquigarrow h(\theta) = \sum_{i=1}^n (Y_i - {}^t\theta x_i)^2$. On cherche les points critiques de $h(\theta)$ i.e. les points tels que $\partial_{\theta_j} h(\hat{\theta}_n^{mc}) = 0$. On écrit donc :

$$\begin{aligned} \sum x_i (Y_i - {}^t\hat{\theta}_n^{mc} x_i) &\stackrel{!}{=} 0 \\ \Rightarrow ({}^tMM)^{-1} ({}^tMM) \hat{\theta}_n^{mc} &= ({}^tMM)^{-1} {}^tMY \\ \Rightarrow \hat{\theta}_n^{mc} &= ({}^tMM)^{-1} {}^tMY \end{aligned}$$

■

R À ce stade, comme pour la RLS, on a pas besoin d'imposer l'hypothèse de normalité sur le bruit.

Propriété 4.2.2 — Absence de biais. Si tMM inversible, que $\mathbb{E}[\xi] = 0$ et que l'hypothèse d'homoscédastricité est respectée ($\mathbb{E}[\xi\xi^t] = \sigma^2 I_n$), alors :

$$\mathbb{E}[\hat{\theta}_n^{mc}] = 0 \text{ (absence de biais !)} \quad (4.10)$$

$$\Sigma(\hat{\theta}_n^{mc}) = \mathbb{E}[(\hat{\theta}_n^{mc} - \theta)(\hat{\theta}_n^{mc} - \theta)^t] = \sigma^2({}^tMM)^{-1} \quad (4.11)$$

Propriété 4.2.3 — Estimation de σ^2 . Si tMM inversible, que $\mathbb{E}[\xi] = 0$ et $\mathbb{E}[\xi\xi^t] = \sigma^2 I_n$, alors :

$$\hat{\sigma}^2 = \frac{\|Y - M\hat{\theta}_n^{mc}\|^2}{n-k} = \frac{1}{n-k} \sum_{i=1}^n (Y_i - \underbrace{{}^t\hat{\theta}_n^{mc}}_{\hat{y}_i})^2 = \frac{1}{n-k} \sum \hat{\xi}_i^2 \quad (4.12)$$

Cet estimateur vérifie $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$.

4.2.2 Loi des estimateurs

Propriété 4.2.4 Si tMM inversible et $\xi_i \sim N(0, \sigma^2)$ i.i.d., alors :

1. $\hat{\theta}_n^{mc} \sim N_k(\theta, \sigma^2({}^tMM)^{-1})$, $\theta = \begin{pmatrix} \theta_0 \\ \vdots \\ \theta - n \end{pmatrix}$
2. $\frac{\|Y - M\hat{\theta}_n^{mc}\|^2}{\sigma^2} = \frac{\hat{\sigma}^2(n-k)}{\sigma^2} \sim \chi^2(n-k)$

Propriété 4.2.5 — Loi de MC $\hat{\sigma}_n^{mc}$ avec variance estimée. Sous les hypothèses d'inversibilité, d'homoscédastricité et de normalité, on a :

$$\forall i \in [1, k], \frac{\hat{\theta}_i - \theta}{\hat{\sigma} \sqrt{({}^tMM)_{ii}^{-1}}} \sim St_{\alpha}(n-k) \quad (4.13)$$

Démonstration. On sait que $\frac{\hat{\theta}_i - \theta_i}{\sigma \sqrt{({}^tMM)_{ii}^{-1}}} \sim N(0, 1)$ et que $\frac{\hat{\sigma}^2}{\sigma^2}(n-k) \sim \chi^2(n-p)$.

Alors $\frac{\hat{\theta}_i - \theta_i}{\hat{\sigma} \sqrt{({}^tMM)_{ii}^{-1}}} * \left(\frac{\hat{\sigma}^2}{\sigma^2}(n-k)\right)^{-1/2} \sim St(n-k)$.

Après simplification, on obtient le résultat voulu. ■

4.2.3 Région de test critique

Maintenant, on peut définir une région de test critique :

Définition 4.2.3

$$R_{\alpha} = \left\{ \left| \frac{\hat{\theta}_j^{mc} - a}{\hat{\sigma} \sqrt{({}^tMM)_{jj}^{-1}}} \right| > q_{1-\alpha/2}(n-k) \right\} \quad (4.14)$$

Cette région de test critique est cruciale pour l'étude des tests, cf. TP.

R Si $Pr(|t|) < 2.10^{-16}$, alors $Pr(|t|) = 0$, donc on rejette H_0 . ($Pr(|t|)$ est $\pm \sim$ à une gaussienne)