

TD 1 : Rappels

Exercice 1 :

1. Rappeler les implications entre les modes de convergence suivants : presque sure, en probabilité, \mathbb{L}^1 , \mathbb{L}^2 , en loi.
2. On définit une suite de variables aléatoires $(X_n)_{n \in \mathbb{N}}$ par $X_n = n^{\frac{1}{2}}$ si $U \leq \frac{1}{n}$ et $X_n = 0$ sinon, où U suit une loi uniforme sur $[0, 1]$. Etudier les différents modes de convergence de X_n .

Exercice 2 : Soit $(X_i)_{i \in \mathbb{N}}$ une suite de variables aléatoires telles que $\mathbb{P}(X_n = 1 - 1/n) = \mathbb{P}(X_n = 1 + 1/n) = 1/2$. On désigne par X la variable aléatoire p.s. égale à 1.

1. Etudier la convergence en loi de X_n .
2. A-t-on $\lim_{n \rightarrow \infty} \mathbb{P}(X_n = x) = \mathbb{P}(X = x)$ pour tout x réel ?
3. Etudier la convergence en probabilité de X_n .
4. Etudier la convergence dans \mathbb{L}^2 de X_n .
5. Etudier la convergence presque sure de X_n .

Exercice 3 : Soit $(X_i)_{i \in \mathbb{N}}$ une suite de variables aléatoires indépendantes et identiquement distribuées (iid) de fonction de répartition (fdr) F .

1. Quelle est la loi de $1_{]-\infty, x]}(X_i)$? En déduire celle de $\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{]-\infty, x]}(X_i)$. \hat{F}_n est appelée *fonction de répartition empirique*. Quel type de fonction est $\hat{F}_n(x)$? Quels sont ses points de discontinuité ? Montrer que, pour tout x , $\lim_{n \rightarrow \infty} \hat{F}_n(x) = F(x)$ p.s.
2. Supposons F continue. Soit $\varepsilon > 0$ tel que $N = 1/\varepsilon$ soit un entier.
 - (a) Montrer qu'il existe une suite $z_0 = -\infty < z_1 < \dots < z_{N-1} < z_N = +\infty$ (dépendant de ε) telle que $F(z_k) = \frac{k}{N}$, $k = 0, \dots, N$.
 - (b) Montrer que pour tout élément de $[z_k, z_{k+1}]$, $\hat{F}_n(x) - F(x) \leq \hat{F}_n(z_{k+1}) - F(z_{k+1}) + \varepsilon$ et $\hat{F}_n(x) - F(x) \geq \hat{F}_n(z_k) - F(z_k) - \varepsilon$.
 - (c) En déduire que $\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \rightarrow 0$ p.s.

Exercice 4 : Soient X_1, \dots, X_n variables aléatoires i.i.d. et \hat{F}_n la f.d.r. empirique correspondante. Soient x, y deux réels. Calculer $\text{Cov}(\hat{F}_n(x), \hat{F}_n(y))$.

Exercice 5 : Soient X_1, \dots, X_n i.i.d. de f.d.r. F et \hat{F}_n la f.d.r. empirique correspondante. On fixe $a < b$ et on considère la fonctionnelle $T(F) = F(b) - F(a)$. Calculer la fonction d'influence de T . On estime $T(F)$ par $\hat{T}_n = T(\hat{F}_n) = \hat{F}_n(b) - \hat{F}_n(a)$. Donner un estimateur de l'écart-type de \hat{T}_n . Donner un intervalle de confiance pour $T(F)$ au niveau asymptotique $1 - \alpha$.

Exercice 6 : Soit $Z = (X, Y)$ une v.a. de f.d.r. F et la fonctionnelle $T(F) = \mathbb{E}\{(X - \mu_X)(Y - \mu_Y)\} / (\sigma_X \sigma_Y)$ (corrélation entre X et Y). Calculer la fonction d'influence de T .

Exercice 7 : Soit F une f.d.r. strictement croissante, de densité f . Soit $T(F) = F^{-1}(p)$ le p -ième quantile. Calculer la fonction d'influence de T .

Exercice 8 : Soit X une variable aléatoire réelle, absolument continue de densité f , de fonction de répartition F , dont on possède un n -échantillon indépendant (X_1, \dots, X_n) . On considère la statistique $T : \mathbb{R}^n \rightarrow \mathbb{R}^n$ qui ordonne l'échantillon dans le sens croissant :

$$(X_1, \dots, X_n) \xrightarrow{T} (X_{(1)}, \dots, X_{(n)}),$$

avec $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$. $(X_{(1)}, \dots, X_{(n)})$ s'appelle la *statistique d'ordre*.

1. Montrer que dans la définition de la statistique d'ordre, on peut se limiter à des inégalités strictes : $X_{(1)} < X_{(2)} < \dots < X_{(n)}$.
2. Déterminer la loi du n -uplet $(X_{(1)}, \dots, X_{(n)})$.
3. Déterminer la fonction de répartition F_k et la densité f_k de $X_{(k)}$. Montrer que si $\mathbb{E}|X|$ est finie, alors il en est de même de $\mathbb{E}|X_{(k)}|$.
4. Rappeler les lois de $X_{(1)}$ et $X_{(n)}$, et déterminer la fonction de répartition du couple $(X_{(1)}, X_{(n)})$. Quelle est la loi de l'étendue $W = X_{(n)} - X_{(1)}$? Vers quoi converge W lorsque n tend vers l'infini?

TD 2 : Tests non paramétriques

Exercice 1 : (Traitement migraines) On veut tester l'efficacité d'un nouveau médicament contre les migraines. On dispose d'un échantillon de 18 personnes sujettes aux migraines à qui on fournit une quantité égale de pillules correspondant au nouveau traitement (A) et de pillules d'aspirine (B). On demande aux patients de choisir une des deux pillules lors d'une crise et lorsqu'ils ont utilisé toutes les pillules, on leur demande de juger quel type de pillule (A ou B) a été le plus efficace. Sur les 18 patients, 12 déclarent que le nouveau traitement (A) est plus efficace que l'ancien (B). Qu'en pensez-vous ?

Exercice 2 : (Points de suture et pansements) On veut comparer les taux de résistances de plaies soignées soit par un pansement, soit par des points de suture. On obtient les données suivantes sur 10 souris, 40 jours après que des incisions aient été faites sur leur dos, traitées les unes avec un pansement et les autres avec des points de suture.

Souris	1	2	3	4	5	6	7	8	9	10
Pansement	659	984	397	574	447	479	676	761	647	577
Points de suture	452	587	460	787	351	277	234	516	577	513

Qu'en pensez-vous ?

Exercice 3 : (Souris infectées par des larves) On s'intéresse à l'effet d'une dose faible de Cambendazole sur les infections des souris par la *Trichinella Spiralis*. Seize souris ont été infectées par un même nombre de larves de *Trichinella* et ensuite réparties au hasard entre deux groupes. Le premier groupe de huit souris a reçu du Cambendazole, à raison de 10 mg par kilo, 60 heures après l'infection. Les autres souris n'ont pas reçu de traitement. Au bout d'une semaine, toutes les souris ont été sacrifiées et le nombre suivant de vers adultes ont été retrouvés dans les intestins :

souris non traitées	51	55	62	63	68	71	75	79
souris traitées	44	47	49	53	57	60	62	67

Que peut-on conclure au sujet d'une éventuelle efficacité du Cambendazole (dosé à 10mg/kilo) pour le traitement des infections des souris par la *Trichinella Spiralis* ?

TD 3 : Estimation d'une densité

Exercice 1 : Montrer que tout noyau pair K est un noyau d'ordre (au moins) 1, dès que la fonction $u \mapsto uK(u)$ est intégrable. Soit le noyau de Silverman

$$K(u) = \frac{1}{2} \exp\left(-\frac{|u|}{\sqrt{2}}\right) \sin\left(\frac{|u|}{\sqrt{2}} + \frac{\pi}{4}\right)$$

On cherche à déterminer son ordre maximal.

1) Montrer que la transformée de Fourier K^* de K vaut

$$K^*(x) = \frac{1}{1+x^4}$$

2) Introduire la fonction $\phi(t) = \int e^{-itu} u^2 K(u) du$ et montrer que $\phi(t) = -(K^*)^{(2)}(t)$.

3) Quel est l'ordre maximal de K ?

Exercice 2 : Soit f une densité qui appartient à la classe de Hölder $\Sigma(\beta, L)$ et X_1, \dots, X_n un n -échantillon de variables aléatoires de densité f . Pour tout entier $s < \beta$, on peut définir un estimateur de la dérivée $f^{(s)}$ via

$$\hat{f}_{n,s}(x) = \frac{1}{nh^{s+1}} \sum_{i=1}^n K_s\left(\frac{X_i - x}{h}\right),$$

où K_s est une application bornée, à support sur $[-1; 1]$ vérifiant pour $\ell = \lfloor \beta \rfloor$, et pour tout $j \in \{0, 1, \dots, \ell\} \setminus \{s\}$,

$$\int u^j K_s(u) du = 0, \quad \int u^s K_s(u) du = s!, \quad (1)$$

1) Montrer que pour tout $x_0 \in \mathbb{R}$ fixé, il existe deux constantes $C_B, C_V > 0$ telles que pour tout $f \in \Sigma(\beta, L)$, et pour tout $h > 0$, on a

$$B_f(\hat{f}_{n,s}(x_0)) := |\mathbb{E}_f \hat{f}_{n,s}(x_0) - f^{(s)}(x_0)| \leq C_B h^{\beta-s}$$

$$\text{Var}_f(\hat{f}_{n,s}(x_0)) := \mathbb{E}_f(|\hat{f}_{n,s}(x_0) - \mathbb{E}_f \hat{f}_{n,s}(x_0)|^2) \leq \frac{C_V}{nh^{2s+1}}.$$

2) Montrer que si $\beta \geq \ell + 1/2$, alors en choisissant la fenêtre h de façon optimale, on obtient un risque quadratique (MSE) maximal de $\hat{f}_{n,s}(x_0)$ sur $\Sigma(\beta, L)$, de l'ordre

de $O(n^{-2(\beta-s)/(2\beta+1)})$ quand $n \rightarrow \infty$.

3) Soit $\{\varphi_m\}_{m \geq 0}$ la base orthonormée de Legendre sur $[-1; 1]$. Montrer que l'application

$$K_s(u) = \sum_{m=0}^{\ell} \varphi_m^{(s)}(0) \varphi_m(u) 1_{|u| \leq 1},$$

vérifie les conditions (1).

Exercice 3 : Soit f une densité de $\mathbb{L}_2(\mathbb{R})$ et X_1, \dots, X_n un n -échantillon de variables aléatoires de densité f .

1) Montrer que l'estimateur à noyau

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right),$$

où K est un noyau de carré intégrable ($\int K^2 < \infty$) vérifie

$$V_f(\hat{f}_n) := \mathbb{E}_f(\|\hat{f}_n - \mathbb{E}_f \hat{f}_n\|_2^2) \leq \frac{\int K^2(u) du}{nh}.$$

2) Supposons de plus que f appartient à la classe de Nikol'ski $\mathcal{N}(\beta, L)$ où $\beta, L > 0$, i.e pour $\ell = \lfloor \beta \rfloor$, on a pour tout $t \in \mathbb{R}$,

$$\|f^{(\ell)}(\cdot + t) - f^{(\ell)}\|_2 = \left(\int \left(f^{(\ell)}(x+t) - f^{(\ell)}(x) \right)^2 dx \right)^{1/2} \leq L|t|^{\beta-\ell}.$$

Alors, pour tout noyau K d'ordre $\ell = \lfloor \beta \rfloor$, tel que $\int |u|^\beta |K(u)| du < \infty$, pour tout $h > 0$, on a

$$B_f^2(\hat{f}_n) := \|\mathbb{E}_f \hat{f}_n - f\|_2^2 \leq \left(\frac{L}{\ell!} \int |u|^\beta |K(u)| du \right)^2 h^{2\beta}.$$

Indication : On pourra utiliser l'inégalité de Minkowski généralisée : Pour toute fonction mesurable g sur $\mathbb{R} \times \mathbb{R}$, on a

$$\int \left(\int g(u, x) du \right)^2 dx \leq \left[\int \left(\int g^2(u, x) dx \right)^{1/2} du \right]^2$$

3) Choisissez la fenêtre optimale $h > 0$ qui minimise le risque quadratique intégré (MISE) maximal de \hat{f}_n sur la classe $\mathcal{N}(\beta, L)$ et donnez la vitesse de convergence de \hat{f}_n sur cette classe.

TD 4 : Régression nonparamétrique (à effets fixes)

Dans tous les exercices suivants, on considère le problème de la régression à effets fixes sur l'intervalle $[0; 1]$:

$$Y_i = r(x_i) + \varepsilon_i, \quad 1 \leq i \leq n, \quad x_i \in [0; 1], \quad \{\varepsilon_i\} \text{ i.i.d. centrées, de variance } \sigma^2.$$

Exercice 1 : Soit $x^* \in [0; 1]$ est point qui n'est pas l'une des variables explicatives x_i . Montrer qu'il n'existe pas d'estimateur linéaire $\hat{r}_n(x^*)$ qui soit sans biais pour toute fonction de régression r .

Exercice 2 : (Régressogramme). Soit $m \geq 1$ un entier. On partitionne l'intervalle $[0; 1]$ en intervalles $I_k = [(k-1)/m; k/m[$, pour $k = 1, \dots, m-1$ et $I_m = [(m-1)/m; 1]$. Le régressogramme est l'estimateur linéaire défini par

$$\hat{r}_{n,m}(x) = \sum_{i=1}^n K(x, x_i, m) Y_i, \quad \text{où } K(x, x_i, m) = \frac{\sum_{k=1}^m 1_{I_k}(x) 1_{I_k}(x_i)}{\sum_{j=1}^n \sum_{k=1}^m 1_{I_k}(x) 1_{I_k}(x_j)}$$

- 1) À quoi correspond cet estimateur ?
- 2) Montrer que si $x \in I_k$, alors

$$\text{Var}(\hat{r}_{n,m}(x)) = \frac{\sigma^2}{n_k},$$

où σ^2 est la variance du terme d'erreur ε et $n_k = \sum_{i=1}^n 1_{x_i \in I_k}$ le nombre de variables explicatives dans l'intervalle I_k .

- 3) On considère le risque quadratique moyen (sur les variables explicatives) défini par

$$R(\hat{r}_{n,m}) = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \{ (\hat{r}_{n,m}(x_i) - r(x_i))^2 \}.$$

Montrer la décomposition biais-variance :

$$R(\hat{r}_{n,m}) = \frac{1}{n} \sum_{i=1}^n \{ \mathbb{E}[\hat{r}_{n,m}(x_i)] - r(x_i) \}^2 + \frac{1}{n} \sum_{i=1}^n \text{Var}(\hat{r}_{n,m}(x_i)).$$

4) (terme de variance). Montrer que

$$\frac{1}{n} \sum_{i=1}^n \text{Var}(\hat{r}_{n,m}(x_i)) = \frac{m\sigma^2}{n}.$$

Dans toute la suite, on se place dans le cas d'un design uniforme.

5) (terme de biais). Montrer que pour m fixé, on a

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (\mathbb{E}[\hat{r}_{n,m}(x_i)] - r(x_i))^2 = \int_0^1 r^2(t) dt - m \sum_{k=1}^m \left(\int_{I_k} r(t) dt \right)^2.$$

Donner une condition suffisante sur r pour que cette limite soit nulle.

6) (terme de biais). À partir de maintenant, on considère le cas $m \rightarrow \infty$. Soit \mathcal{F} l'ensemble des fonctions de classe $\mathcal{C}^1([0; 1])$. Montrer que si $r \in \mathcal{F}$ alors

$$\frac{1}{n} \sum_{i=1}^n (\mathbb{E}[\hat{r}_{n,m}(x_i)] - r(x_i))^2 \leq \frac{\|r'\|_{\infty, [0;1]}^2}{m^2}.$$

Conclure.

Exercice 3 : (Estimateur à noyau). On suppose que $x_i = (2i - 1)/2n$, pour tout $1 \leq i \leq n$. Soit $K \in \mathcal{C}^1([-1; 1])$ un noyau positif d'ordre 1. On définit l'estimateur suivant

$$\hat{r}_n(x) = \sum_{i=1}^n h^{-1} \left\{ \int_{(i-1)/n}^{i/n} K\left(\frac{s-x}{h}\right) ds \right\} Y_i.$$

On souhaite contrôler le risque de cet estimateur en un point x fixé :

$$R(\hat{r}_n(x)) = \mathbb{E}\{\hat{r}_n(x) - r(x)\}^2.$$

1) Montrer que

$$\begin{aligned} \text{Var}(\hat{r}_n(x)) &= \frac{\sigma^2(\int_0^1 K^2((x-u)/h) du)}{nh^2} + O((nh)^{-2}) \\ &\leq \frac{\sigma^2 \int_{-1}^1 K^2(u) du}{nh} + O((nh)^{-2}). \end{aligned}$$

2) On suppose que la fonction de régression r est de classe \mathcal{C}^2 sur $[0; 1]$. Montrer que

$$\mathbb{E}(\hat{r}_n(x)) = \frac{1}{h} \left(\int_0^1 K((u-x)/h) r(u) du \right) + O(n^{-1}).$$

En déduire

$$\mathbb{E}(\hat{r}_n(x)) - r(x) = \frac{h^2}{2} r''(x) \left(\int_{-1}^1 u^2 K(u) du \right) + o(h^2) + O(n^{-1}).$$

3) Conclure.

TD 5 : Modèle de suites Gaussiennes

Dans tous les exercices suivants, on considère le modèle des suites Gaussiennes où $Y \sim \mathcal{N}_n(\theta, \sigma_n^2 I)$, c'est-à-dire

$$Y_i = \theta_i + \varepsilon_i, \quad 1 \leq i \leq n, \quad \{\varepsilon_i\} \text{ i.i.d. centrées, de variance } \sigma_n^2.$$

Exercice 1 : (Phénomène de Stein). Soit $h : \mathbb{R}^n \rightarrow \mathbb{R}$ une fonction différentiable. On considère les estimateurs de la forme $\hat{\theta}(Y) = h(Y)Y$.

1) En utilisant le théorème de Stein, donner une expression d'un estimateur sans biais du risque quadratique $R_n(\hat{\theta}_n, \theta)$ de $\hat{\theta}$.

2) On considère plus précisément les fonctions h de la forme $h(y) = 1 - c/\|y\|_2^2$. Donner la forme de l'estimateur du risque $R_n(\hat{\theta}_n, \theta)$ et en déduire une expression de ce risque.

Dans la suite, on admet que $\mathbb{E}_\theta(1/\|Y\|_2^2) \in]0; +\infty[$ pour $n \geq 3$ (lorsque $n = 1$ ou 2 , cet espérance vaut $+\infty$).

3) Donner la constante c qui minimise le risque $R_n(\hat{\theta}_n, \theta)$. Que pouvez-vous conclure ?

Exercice 2 : (Maximum de vraisemblance pénalisé). On considère des estimateurs de la forme

$$\hat{\theta}(\lambda) = \operatorname{argmin}_{\theta \in \mathbb{R}^n} \sum_{i=1}^n (Y_i - \theta_i)^2 + \lambda J(\theta),$$

où $\lambda > 0$ est un paramètre et J la fonction de pénalité.

1)[shrinkage]. Soit $J(\theta) = \|\theta\|_2^2$. Montrer que $\hat{\theta}(\lambda) = (1 + \lambda)^{-1}Y$.

2)[seuillage doux]. Soit $J(\theta) = \|\theta\|_1$. Vérifier que l'estimateur défini pour tout $1 \leq i \leq n$ par $\hat{\theta}_i(\lambda) = \operatorname{sgn}(Y_i)(|Y_i| - \lambda/2)_+$ est solution du problème.

3)[seuillage dur]. On considère $J(\theta) = \#\{i; \theta_i \neq 0\}$. Vérifier que l'estimateur défini pour tout $1 \leq i \leq n$ par $\hat{\theta}_i(\lambda) = Y_i 1\{|Y_i| > \sqrt{\lambda}\}$ est solution du problème.

Exercice 3 : (Construction d'intervalles de confiance). Soit $\alpha \in [0; 1]$. On considère

$$\mathcal{B}_n(\alpha) = \{\theta \in \mathbb{R}^n; \|Y - \theta\|_2^2 \leq \sigma_n^2 \chi_{n,\alpha}^2\},$$

où $\chi_{n,\alpha}^2$ est le $(1 - \alpha)$ -quantile d'un χ^2 à n degrés de liberté.

1) Expliquer pourquoi $\mathcal{B}_n(\alpha)$ est une région de confiance pour θ au niveau $1 - \alpha$. Le rayon moyen de cette boule de confiance est $\sqrt{n}\sigma_n$. On veut construire des boules ayant un rayon moyen plus petit. Pour cela, on s'intéresse au test de l'hypothèse $\theta = 0$, au niveau $\alpha/2$, en utilisant la statistique de test $\sum_{i=1}^n Y_i^2$.

2) Proposer un test au niveau $\alpha/2$ de l'hypothèse $H_0 : \theta = 0$ contre $H_1 : \theta \neq 0$. En utilisant une approximation Gaussienne, exhiber un test qui atteint asymptotiquement le niveau $\alpha/2$.

3) Soit $\Delta_n^2 = 2\sqrt{2n}\sigma_n^2 z_{\alpha/2}$, où $z_{\alpha/2}$ est le $(1 - \alpha/2)$ quantile de la loi $\mathcal{N}(0; 1)$. Montrer que sur l'alternative $H_{1,n} : \|\theta\|_2^2 \geq \Delta_n^2$, le test atteint la puissance asymptotique $\alpha/2$. On définit à présent

$$R_n(\alpha) = \begin{cases} \mathcal{B}_n(\alpha/2) & \text{si le test rejette l'hypothèse } H_0 \\ \{\theta; \|\theta\|_2^2 < \Delta_n^2\} & \text{sinon.} \end{cases}$$

4) Montrer que $R_n(\alpha)$ est une région de confiance au niveau asymptotique $1 - \alpha$ pour θ (on pourra distinguer les cas a) $\theta = 0$, b) $\theta \neq 0$ et $\|\theta\|_2^2 < \Delta_n^2$, c) $\theta \neq 0$ et $\|\theta\|_2^2 \geq \Delta_n^2$). Conclure.