Projet L3 MIASHS

HOFF Eugénie
SAKHRAOUI Yasser

MAI 2021

Table des matières

Introduction

Pendant ce projet nous allons aborder 3 chapitres. D'une part il va se composer de deux chapitres théoriques et un dernier consacré à une application à des données réelles. Lors du de la partie théorique nous aborderons, dans un premier temps, l'étude de l'estimation non-paramétrique des modèles à densité où nous parlerons de l'estimation par histogramme, à noyau et de la validation croisée. Dans un second temps, nous présenterons le test non paramétrique de kolmornv-Smirnov d'adéquation et d'homogénéité. Enfin dans le dernier chapitre nous appliquerons les deux premières sections avec des données réelles.

Chapitre 1

Estimation non paramétrique d'une densité de probabilité

1.1 Introduction

L'estimation non paramétrique a pour objectif d'estimer la densité de probabilité à partir des informations disponibles. Contrairement à l'estimation paramétrique nous n'avons pas besoin de connaître la loi de probabilité de la densité.

Nous allons voir dans un premier temps, l'estimation non-paramétrique d'une densité par histogramme puis dans un second temps, l'estimation non-paramétrique d'une densité de probabilité par des méthodes à noyau. Afin d'estimer non-paramétriquement la densité de probabilité p en se basant sur les observations $x_1, ..., x_n$, nous utilisons l'estimation par histogramme et à noyau. Mais tout d'abord définissons les termes.

1.1.1 Fonction de répartition

Soit $X = (X_1, ..., X_n)$, un n-échantillon qui suis une loi de probabilité P.
Soit $(x_1, ..., x_n)$, une observation de cet échantillon. La fonction de répartition F est inconnue et elle est définie par : $F(x) = P\{X_1 \leq x\}$

Les observations sont ordonnées :

$$x_1 \leq x_2 \leq ... \leq x_n$$

Nous supposons que F est inconnue. Nous estimons F par une fonction de répartition empirique, F_n. nous la définissons par :

$$\begin{aligned}
F_n(x) &= \frac{\text{nombre d'observations } \leq x}{n} \\
&= \frac{\#\{i : X_i \leq x\}}{n} \\
&= \frac{1}{n}\sum_{i=1}^{n} I_{]-\infty,x]}(X_i) \\
&= \begin{cases} 0 & \text{si } x < X_1 \\ \frac{k}{n} & \text{si } X_k \leq x \leq X_{k+1} \\ 1 & \text{si } x \geq X_1 \end{cases}
\end{aligned}$$

Avec :

\# : nombre de données

$I_{]-\infty,x]}$: la fonction indicatrice

$k = 1, ..., n-1$.

1.1.2 Propriétés de $F_n(x)$

Bais

$$\begin{aligned}
E\{F_n(x)\} &= \frac{1}{n}\sum_{i=1}^{n} E[I_{]-\infty,x]}(X_i)] \\
&= \frac{1}{n}\sum_{i=1}^{n} P\{X_i \leq x\} \\
&= P\{X_1 \leq x\} \\
&= F(x)
\end{aligned}$$

Donc $F_n(x)$ est un estimateur sans bais de $F(x)$.

Variance

$$\begin{aligned}
Var\{F_n(x)\} &= E\{\frac{1}{n}\sum_{i=1}^{n} I_{]-\infty,x]}(X_i)\}^2 \\
&= \frac{1}{n^2}E[(\sum_{i=1}^{n} I_{]-\infty,x]}(X_i))(\sum_{i=1}^{n} I_{]-\infty,x]}(X_j))] \\
&= \frac{1}{n^2}E[(I_{]-\infty,x]}(X_i))(I_{]-\infty,x]}(X_j)) + \sum_{i=1}^{n} E[(I_{]-\infty,x]}(X_i))(I_{]-\infty,x]}(X_i))] \\
&= F(x) - F(x)^2 \\
&= F(x)(1 - F(x))
\end{aligned}$$

Loi forte des grands nombres

$$\forall x \in \mathbf{R} : F_n(x) \xrightarrow[n\to+\infty]{p.s} F(x)$$

Théorème centrale limite

on a : $F(x)$ la fonction de répartition de l'échantillon $X_1, .., X_n$ et $F_n(x)$ sa fonction empirique. Dans ce théorème, on nous montre que $F_n(x)$ se rapproche de $F(x)$ selon une loi normale dès lors que n est assez grand. De cette manière, on peut observe la qualité de l'estimateur.

$$\sqrt{n}\frac{F_n(x) - F(x)}{\sqrt{F(x)(1 - F(x))}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$
$$F_n(x) - F(x) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \frac{Var(F_n(x))^2}{n})$$

Donc notre $F(x) - F_n(x)$ suit une gaussienne centrée en 0 avec une variance qui décroît significativement avec n.

1.2 Estimation non-paramétrique par histogramme

L'utilisation de l'estimateur non-paramétrique d'une densité de probabilité par histogramme est couramment utilisé car sa représentation graphique est très visuel. Elle fonctionne très bien en faible dimensionalité, en 2 voire 3 dimensions. L'histogramme va nous permettre d'étudier la répartition des données. Cette représentation est possible si la loi F des X_i est continue et si elle admet une densité de probabilité p. Cependant, l'estimation des densité sont des fonctions étagées, cela signifie que toutes les fonctions ne sont pas représentables.

Nous déterminons un intervalle $A = [a, b]$ et D, le nombre de cases (bâtonnets de l'histogramme). Nous avons des intervalles de même largeur.

$$A_{k,D} = [a + (k - 1)\frac{b - a}{D}, a + k\frac{b - a}{D}[; k = 1, ..., D]$$

1.2.1 Estimateur par histogramme

Nous comptons le nombre d'observations parmi les $x_1, ..., x_n$ dans chaque intervalle. Nous posons : $h = \frac{b-a}{D}$, h correspond à la largeur de fenêtre et au paramètre de lissage.

L'estimateur par histogramme est :

$$\begin{aligned}\hat{p}_n &= \frac{1}{h}\sum_{k=1}^{D}(\frac{1}{n}\sum_{i=1}^{n} I_{\{x_i \in A_{k,D}\}})I_{\{A_{k,D}\}}(x) \\ &= \frac{1}{nh}\sum_{k=1}^{D} N_k I_{\{A_{k,D}\}}(x) \\ &= \frac{1}{nh}\#\{x_i \in \{A_{k,D}\}\}\end{aligned}$$

Plus l'intervalle choisit est petit (donc un grand nombre de D)et plus le nombre de n est important, plus l'approximation sera de qualité.
Maintenant, nous choisissons, la position des intervalles ainsi que le paramètre de lissage h, nous nous appuyons sur l'erreur quadratique moyenne, le MSE et sur l'erreur quadratique moyenne intégrée, le MISE.

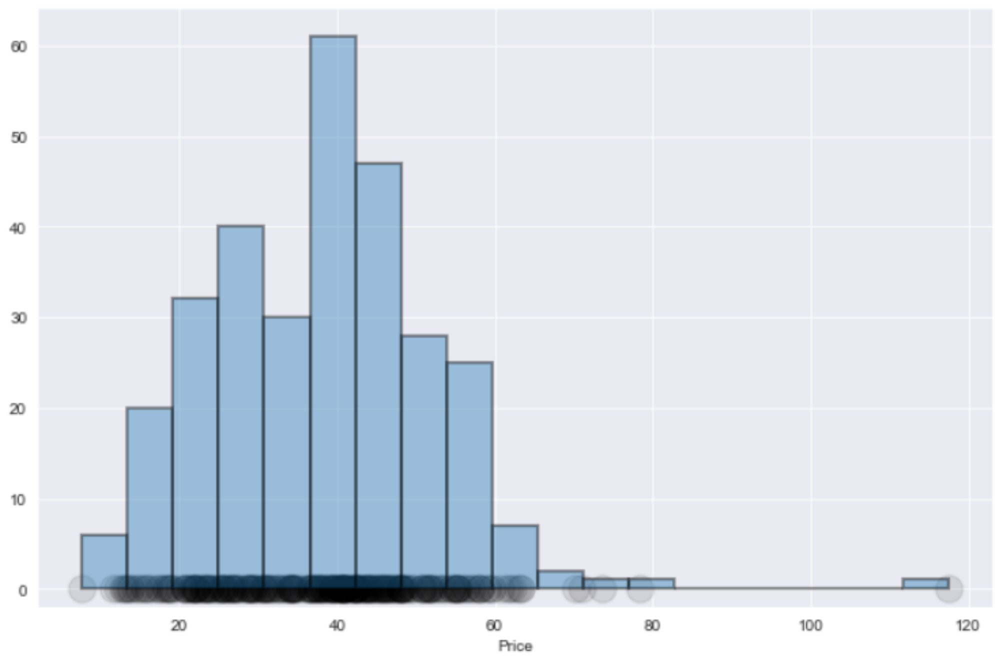

FIGURE 1.1 – Estimation non-paramétrique par histogramme

1.2.2 Erreur quadratique moyenne (MSE)

Avec E_p, l'espérance par rapport à la loi de $(X_1, ..., X_n)$. Nous avons :

$$\begin{aligned} MSE(x_0) &= MSE \\ &= b^2(x_0) + \sigma^2(x_0) \end{aligned}$$

Avec : $b(x_0)$ et $\sigma^2(x_0)$, qui sont respectivement le biais et la variance de $\hat{p}_n$ au point x_0.

- $b(x_0) = E_p[\hat{p}_n(x_0) - p(x_0)]$
- $\sigma^2(x_0) = E_p[(\hat{p}_n(x_0) - E_p[\hat{p}_n(x_0)])^2]$

Theorem 1 *Soit p la fonction de densité des v.a i.i.d $X_1, ..., X_n$ définie par un intervalle $\{A_{k,D}\}$ et $\hat{p}$ est l'estimateur par histogramme de p. On a $p^{'} = P(X_1 \in \{A_{k,D}\}\})$*
Soit un réel $L > 0$, on dit que p est L-Lipschitzienne sur $\{A_{k,D}\}$, si $\forall p,p^{'} \in \{A_{k,D}\}$:

$$|p(x) - p(y)| \leq L(x - y)$$

Dans notre cas, L = 1 donc

$$|p(x) - p(y)| \leq (x - y)$$

Proposition 1 *Soit p, la densité des v.a i.i.d $X_1, ..., X_n$ et $\hat{p}_n$ l'estimateur par histogramme de p. Soit $x_0 \in \{A_{k,D}\}\}$, nous notons $p' = P(X_1 \in \{A_{k,D}\}\})$*

$$\begin{aligned} MSE = MSE(x_0) &= E_p[\hat{p}_n(x_0) - p(x_0))^2] \\ &= (\ \frac{p^{'}}{h} - p(x_0))^2 + \frac{p^{'}(1-p^{'})}{nh^2} \end{aligned}$$

En majorant le MSE, nous obtenons :

$$MSE = h^2 + \frac{1}{nh^2}$$

Avec le minimum en h qui est :

$$h_n^* = \mathcal{O}(n^{\frac{-1}{4}})$$

Quand $n \to \infty$ et en prenant $h = h_n^*$, nous obtenons :

$$MSE = \mathcal{O}(n^{\frac{-1}{2}})$$

1.2.3 Erreur quadratique moyenne intégrée (MISE)

En majorant le MISE, nous obtenons :

$$MISE = h^2 + \frac{1}{nh}$$

Avec le minimum en h qui est :

$$h_n^* = \mathcal{O}(n^{\frac{-1}{3}})$$

Quand $n \to \infty$ et en prenant $h = h_n^*$, nous obtenons :

$$MISE = \mathcal{O}(n^{\frac{-2}{3}})$$

Nous constatons que le choix du paramètre de lissage est important dans l'estimation par histogramme. Effectivement, plus le paramètre de lissage sera petit, plus l'histogramme sera découpé (moins de cases) et plus le paramètre de lissage sera grand, plus l'histogramme sera lissé (grand nombre de D).

L'estimateur par histogramme dépend de deux paramètres : le point d'origine et le paramètre de lissage h. Ces deux paramètres ont donc une forte influence sur lui.

L'estimation par la méthode des noyaux va nous permettre d'avoir une estimation non paramétrique avec une densité plus lisse et d'éviter les sauts entre chaque classe qu'on retrouve dans l'estimation par histogramme. Avec cette méthode, la fonction dépendra du paramètre de lissage h

1.3 Estimation non-paramétrique par noyau

Pour $h > 0$ et h assez petit, nous avons :

$$p(x) \approx \frac{F(x+h) - F(x-h)}{2h}$$

Nous remplaçons F par son estimateur F_n :

$$p_n(x) \approx \frac{F_n(x+h) - F_n(x-h)}{2h}$$

Avec :

$p_n(x)$: l'estimateur de p

1.3.1 Estimateur à noyau

Nous pouvons généraliser par :

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

Avec :

$K : \mathbf{R} \to \mathbf{R}$

$\int K(u)du = 1$

K : le noyau

h : la fenêtre de l'estimateur et le paramètre de lissage

Propriétés sur K :

- Si K est une densité de probabilité, alors $\hat{p}$ est aussi une densité de probabilité,
- $\hat{p}$ a les mêmes propriétés de continuité et de différentiablilité que K :
  - Si K est continue, alors $\hat{p}$ est une fonction continue,
  - SiK est différentiable, alors $\hat{p}$ est une fonction différentiable,
  - Si K peut prendre des valeurs négatives, alors $\hat{p}$ pourra également prendre des valeurs négatives.

Exemples de différents noyaux :

- noyau gaussien : $K(u) = \frac{1}{\sqrt{2\pi \exp(\frac{-u^2}{2})}}$
- noyau rectangulaire : $K(u) = \frac{1}{2} I_{|u| \leq 1}$
- noyau triangulaire : $K(u) = (1 - |u|) I_{|u \leq 1|}$.

Si $K > 0$ et $X_1, ..., X_n$ est fixé, alors $x \to \hat{p}_n(x)$ est une densité de probabilité.

Nous allons donner des propriétés de l'estimateur à noyau .Nous étudierons sa variance, son biais, son erreur quadratique moyenne (MSE) et son erreur quadratique moyenne intégrée (MISE).

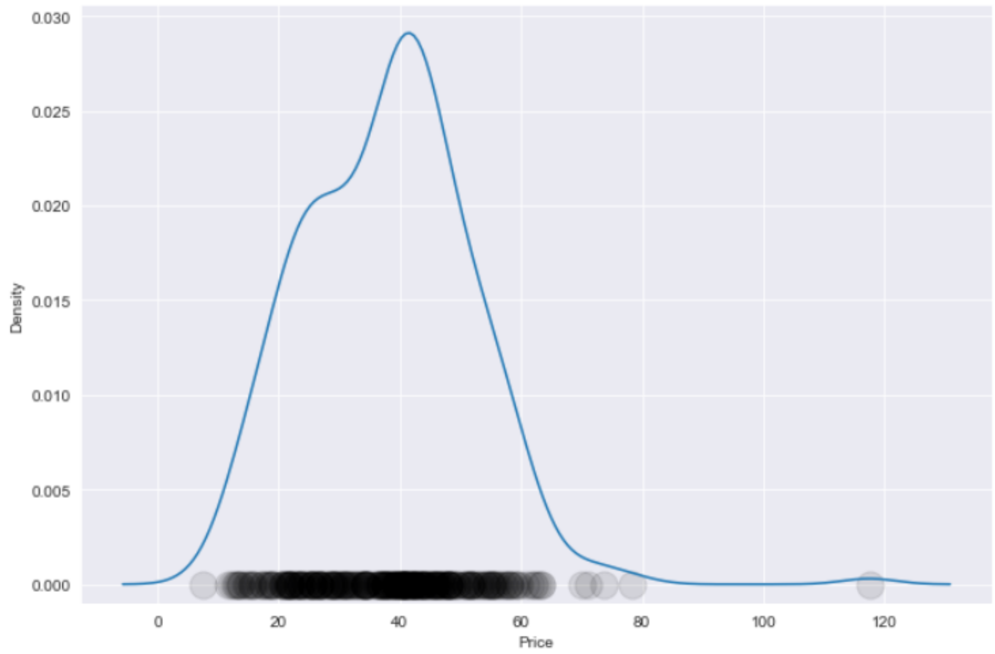

FIGURE 1.2 – Estimation non-paramétrique à noyau

1.3.2 Erreur quadratique moyenne (MSE)

Nous avons un échantillon $X_1, ..., X_n$ issu d'une $v.aX$ avec comme fonction de densité p que l'on souhaite estimer.
$\hat{p}_n$ est l'estimateur à noyau avec comme noyau K et comme paramètre de lissage h.

Étude de la variance :

Soit $x_0 \in \mathbf{R}, x_0$ est fixé. Nous allons contrôler la variance de $\hat{p}_n$ au point x_0.

Proposition 2 *Nous supposons que la densité de probabilité p vérifie $p(x) \leq p_{max} < \infty$, pour tout $x_0 \in \mathbf{R}$ et que K est :*

$$\int K(u)du = 1, \int K^2(u)du < \infty$$

Alors, $\forall x \in \mathbf{R}$, $h > 0$ et $n \geq 1$,

$$\sigma^2(x_0) \leq \frac{C_1}{nh} \tag{1.1}$$

Avec $C_1 = p_{max} \int K^2(u)du$

Si en plus, p est continue et que $p(x_0) > 0$, alors, $\sigma^2(x_0) = \frac{p(x_0)}{nh} \int K^2 du(1 + \mathcal{O}(1))$
Avec : $\mathcal{O}(1)$ qui dépend de $p_(x_0)$, K et h.

Si $h = h_n$ avec $nh \to \infty$ quand $n \to \infty$ alors $\sigma^2(x_0) \to 0$.

Étude du biais :

Soit $x_0 \in \mathbf{R}, x_0$ est fixé. Nous allons contrôler le biais de $\hat{p}_n$ au point x_0.

Nous avons :

$$b(x_0) = E_n[\hat{p}_n(x_0)] - p(x_0) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{X_i - x_0}{h})$$

Nous appliquons l'intégrale :

$$\begin{aligned} b(x_0) &= \frac{1}{h} \int K(\frac{z - x_0}{h}) p(z) dz - p(x_0) \\ &= \int K(u)[p(x_0 + uh) - p(x_0)] du \end{aligned}$$

Nous supposons que $p \in \mathcal{J}$, avec $\mathcal{J}$, la classe de densité $\mathcal{J} = \mathcal{J}(\beta, L)$ définie par :

$$\mathcal{J}(\beta, L) = p(p \geq 0, \int p = 1, p \in \sum(\beta, L) \text{sur} \mathbf{R}$$

Definition 1 Soit T un intervalle de $\mathbf{R}$ et soient $\beta > 0$, $L > 0$. La classe de Hölder $\sum(\beta, L)$ sur T est définie comme l'ensemble des fonctions $f : T \to \mathbf{R}$ telles que la dérivé de $f^{(l)}, l = \lfloor \beta \rfloor$, existe et vérifie :

$$|f^{(l)}(x) - f^{(l)}(x^{'})| \leq L|x - x'|^{\beta - l}, \quad \forall x, x' \in T$$

Soit K, un noyau d'ordre l.

Proposition 3 *Soit $p \in \mathcal{J}(\beta, L)$ et K noyau d'ordre $l = \lfloor \beta \rfloor$. Tel que :*

$$\int |u|^{\beta} |K(u)| du < \infty$$

Alors, $\forall x_0 \in \mathbf{R}$, $h > 0$ et $n \geq 1$:

$$|b(x_0)| \leq C_2 h^{\beta} \tag{1.2}$$

Avec, $C_2 = \frac{L}{l!} \int |u|^{\beta} |K(u)| du$

Erreur quadratique moyenne (MSE) :

Soit $x_o \in \mathbf{R}, x_0$ est fixé. Nous allons définir le MSE, soit le risque quadratique moyen de $\hat{p}_n$ au point x_0.

$$MSE = MSE(x_0) = E_p[(\hat{p}_n(x_0) - p(x))^2]$$

Avec E_p, l'espérance par rapport à la loi de $(X_1, ..., X_n)$. Nous avons :

$$MSE = b^2(x_0) + \sigma^2(x_0)$$

Avec : $b(x_0)$ et $\sigma^2(x_0)$, sont respectivement le biais et la variance de $\hat{p}_n$ au point x_0.

— $b(x_0) = E_p[\hat{p}_n(x_0) - p(x_0)]$
— $\sigma^2(x_0) = E_p[(\hat{p}_n(x_0) - E_p[\hat{p}_n(x_0)])^2]$

Afin de mesurer la performance de l'estimateur à noyau, nous nous nous appuyons sur l'étude de la variance puis le biais de cet estimateur $\hat{p}_n$ que nous avons fait précédemment.

Nous allons majorer le MSE afin de trouver un compromis entre la variance et le biais de l'estimateur à noyau.
Si p et K vérifient les équations (1.1) et (1.2) , nous avons :

$$MSE \leq C_2^2 h^{2\beta} + \frac{C_1}{nh}$$

Avec le minimum en h qui est :

$$h_n^* = (\frac{C_1)}{2\beta C_2^2})^{\frac{1}{2\beta+1}} n^{\frac{-1}{2\beta+1}}$$

Quand $n \to \infty$ et en prenant $h = h_n^*$, nous obtenons :

$$MSE = \mathcal{O}(n^{\frac{-2\beta}{2\beta+1}})$$

Theorem 2 *Nous avons $p \in \mathcal{J}(\beta, L)$, K est le noyau d'ordre $l = \lfloor \beta \rfloor$ tel que :*

$$\int |u|^\beta |K(u)| du < \infty \text{et} \int K^2(u) du < \infty$$

Nous supposons que p est borné et nous fixons : $h = \alpha^{\frac{-1}{2\beta+1}}, \alpha > 0$, alors pour tout $n \leq 1$:

$$\underset{x_0 \in \mathbf{R}}{sup}, \underset{p \in \mathcal{J}(\beta,L)}{sup}, E[(\hat{p}_n(x_0) - p_n(x_0))^2] \leq C_n^{\frac{-2\beta}{2\beta+1}}$$

Nous avons la vitesse de convergence ω_n de l'estimateur à noyau $\hat{p}_n(x_0)$.

$$\omega_n = n^{\frac{-\beta}{2\beta+1}}$$

Afin de savoir si nous pouvons améliorer ω_n avec d'autres estimateurs de densité ou si nous évaluons la meilleur vitesse de convergence, nous définissons $\mathcal{R}_n^*$, le risque minimax associé à la classe $\mathcal{J}(\beta, L)$:

$$\mathcal{R}_n^* \mathcal{J}(\beta, L) = \underset{T_n}{inf} \underset{p \in \mathcal{J}(\beta,L)}{sup} E[(T_n(x_0) - p(x_0))^2]$$

Avec :

T_n : l'ensemble de tous les estimateur à noyau.

Quand le minimum est pris en charge par tous les estimateurs, nous avons :

$$\mathcal{R}_n^* \mathcal{J}((\beta, L)) \geq C\omega_n^2$$

L'estimateur à noyau atteint la vitesse optimal de convergence en : $\omega_n = n^{\frac{-\beta}{2\beta+1}}$

1.3.3 Erreur quadratique moyenne intégrée (MISE)

Afin de mesurer globalement la précision de $\hat{p}_n$ comme étant l'estimateur de p en tout point x, nous allons étudier l'erreur quadratique moyenne intégrée (MISE) au point arbitraire x.

Nous avons :

$$MISE = E_p \int (\hat{p}_n(x) - p(x))^2 dx$$

Avec le théorème de Tonelli-Fubini, nous avons :

$$\begin{aligned} MISE &= \int MSE(x)dx \\ &= \int b^2(x)dx + \int \sigma^2(x)dx \end{aligned}$$

Le MISE est,comme le MSE, composé de la variance et du biais. nous agissons avec le même processus, utilisé lors du MSE : nous analysons la variance puis le biais.

Étude de la variance

Proposition 4 *Nous supposons que K est un noyau tel que :*

$$\int K(u)du = 1, \int K^2(u)du < \infty$$

Alors $\forall h > 0$ et $\forall n \geq 1$ de la densité de probabilité p, nous avons :

$$\int \sigma^2(x)dx \leq \frac{1}{nh} \int K^2(u)du \tag{1.3}$$

Étude du biais

Definition 2 Soit $\beta > 0$ et $L > 0$, la classe de Nikol'Ski, $\mathcal{N}(\beta, L) = f : \mathbf{R} \to \mathbf{R}, f\lfloor\beta\rfloor$ existe et satisfait :

$$[\int (f^{\lfloor\beta\rfloor}\ (x+t) - f^{\lfloor\beta\rfloor}))^2\ dx]^{\frac{1}{2}} \leq L\ |t|^{\beta - l}, \quad \forall t \in \mathbf{R}.$$

Proposition 5 *Soit p une densité dans $\mathcal{N}(\beta, L)$ et K un noyau d'ordre $l = \lfloor \beta \rfloor$ qui satisfait :*

$$\int |u|^{\beta} \ |K(u)| \ du < \infty$$

Alors, $\forall h > 0$ et $n \geq 1$, nous avons :

$$\int b^2(x)dx \leq C_2^2 \ h^{2\beta} \tag{1.4}$$

Avec,
$b^2(x) = (E[\hat{p}_n(x)] - f(x))^2$
$C_2 = \frac{L}{l!} \int |u]^{\beta} |L(u)|du$

MISE

Nous utilisons l'équation (1.3) et l'équation (1.4) et nous obtenons :

$$MISE \leq C_2^2 \ h^{2\beta} + \frac{\int K^2}{nh}$$

Avec le minimum en h qui est :

$$h_n^* = (\frac{\int K^2}{2\beta C_2^2})^{\frac{1}{2\beta+1}} \ n^{\frac{-1}{2\beta+1}}$$

Quand $n \to \infty$ et en prenant $h = h^*$, nous obtenons :

$$MISE = \mathcal{O}(n^{\frac{-1}{2\beta+1}})$$

1.3.4 Validation croisée

La validation croisée, nous a été introduite par Rudemo (1982) et par Bowar (1984). La validation croisée est une méthode d'estimation de fiabilité d'un modèle fondée sur un échantillon. Cette méthode va nous permettre de choisir la fenêtre h dans le but de limiter le sur ou sous lissage de h quand le noyau K est fixe. Ici, nous supposons que le noyau K est fixé, afin de s'intéresser au choix de la fenêtre h

La valeur idéal de h nous ai donné par l'équation :

$$MISE = MISE(h)$$
$$h_{id} = arg \quad \underset{h>0}{min} \quad MISE(h)$$

Cependant, nous ne connaissons pas la valeur de la densité p. Nous allons donc minimiser le MISE de l'estimateur à noyau sur un intervalle fini de h.

$$MISE(h) = E_p \int (\hat{p}_n - p)^2 = E_p \ [\ \hat{p}_n^2 - 2 \int \hat{p}_n p\ ] + \int p^2$$

Comme "$\int p^2$" ne dépend pas de h, cela revient à minimiser la fonction $J(h)$:

$$J(h) = E_p \ [\ \hat{p}_n^2 - 2 \int \hat{p}_n p\ ]$$

Nous ne connaissons pas la valeur de "$E_p[\int \hat{p}_n p]$", car elle dépend de la densité de p qui nous est inconnue. Nous allons donc l'estimer par : $G = E_p[\int \hat{p}_n p]$

avec $\hat{G} = \frac{1}{n} \sum_{i=1}^{n} \hat{p}_{n,-i}(X_i)$

où $\hat{p}_{n,-i}(x) = \frac{1}{(n-1)h} \sum_{j}^{n} \frac{X_i - x}{h}$

Cet estimateur $\hat{G}$ à noyau est tiré d'un échantillon X_i. Dont ses observations sont : $x_1, ..., x_{1-i}, x_{1+i}, ..., x_n$. De plus, l'estimateur $\hat{G}$ est sans biais.

$$\begin{aligned} E_p(\hat{G}) &= E_p \ [\hat{p}_{n,-1} \ (X_1)] \\ &= E_p \ [\frac{1}{(n-1)h} \sum_{j \neq 1} \int K(\frac{X_j - z}{h}) \ p(z) \ dz] \\ &= \frac{1}{h}(x)(z) \int K(\frac{x - z}{h}) \ p(z) \ dz \ dx \end{aligned}$$

d'où :

$$\begin{aligned} G &= E_p \ [\int \hat{p}_n p] \\ &= E_p \ [\frac{1}{nh} \sum_{i=1}^{n} \int K(\frac{X_i - z}{h}) \ p(z) \ dz] \\ &= \frac{1}{h}(x) \int K(\frac{x - z}{h}) \ p(z) \ dz \ dx \end{aligned}$$

D'où : $G = E_p(\hat{G})$

L'estimateur sans biais de $J(h)$ est :

$$CV(h) = \int \hat{p}_n^2 - \frac{2}{n}\sum_{i=1}^{n} \hat{p}_{n,-i}(X_i)$$

Avec $CV(h)$ qui représente la fonction de la validation croisée,
et $E_p[CV(h)] = MISE(h) - \int p^2$

Enfin la valeur de h qui minimise $CV(h)$ est le paramètre de lissage issu de la validation croisée :

$$h_{CV} = \underset{h>0}{argmin} CV(h)$$

Chapitre 2

Test de Kolmogorov Smirnov

2.1 Introduction

Le test d'ajustement de Kolmornov-Smirnov (K-S) est un test non paramétrique, souvent utilisé. Son nom nous vient du mathématicien Andréi Nikoláyevich Kolmogorov qui établit l'axiomatique des probabilités en 1933. Ce test est basé sur les fonctions de répartition à la différence du test d'adéquation du $\tilde{\chi}^2$, qui se fonde sur les densités. Il permet de déterminer si un échantillon suit une loi donnée comme sa fonction de répartition continue, ou bien si deux échantillons suivent la même loi. L'objectif sera de chercher à obtenir une estimation de la fonction de répartition, F à partir de l'échantillon observé afin de la comparer ensuite à la fonction de répartition de la loi théorique. Ou alors, de mesurer l'écart maximal entre deux fonctions de répartitions théorique, ce sera le test de K-S de comparaison de deux échantillons

2.2 Test KS d'adéquation

Soit $X = (X_1, ..., X_n)$, un n-échantillon qui suis une loi de probabilité P.
Soit $(x_1, ..., x_n)$, une observation de cet échantillon. La fonction de répartition F associé à P est inconnue et elle est définie par : $F(x) = P\{X_1 \leq x\}$
Nous estimons F par la fonction de répartition empirique F_n associée à l'échantillon X :

$$F_n(x) = \frac{1}{n}\sum_{k=i}^{n} I_{]-\infty;x]}(X_i)$$

A noter que $\forall x \in R$,F$_n(x)$ une v.a qui a des valeurs $\in [0, 1]$

et que par la loi forte des grands nombres, nous avons : $F_n(x) \xrightarrow[n\to+\infty]{p.s} F(x)$.
Le principe de ce test d'hypothèses est de mesurer l'écart maximal qu'il existe entre une fonction de répartition empirique, F_n et une fonction de répartition théorique F, c'est le test de K-S d'adéquation.

Le test de K-S d'adéquation va comparer la distribution observée d'un échantillon statistiques à une distribution théorique. Il est utile quand le caractère observé peut prendre des valeurs continues.

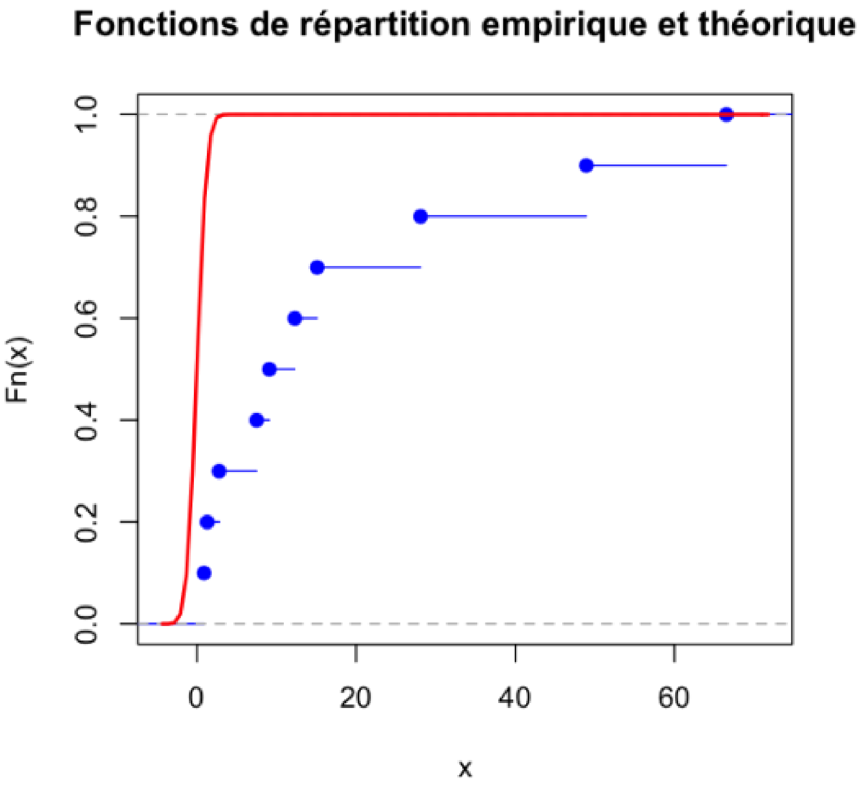

FIGURE 2.1 – Empirical cumulitavie distibution function

2.2.1 Théorème de Gilvenko-Cantelli

Theorem 3 (Théorème de Gilvenko-Cantelli) *Soit $X = (X_1, ..., X_n)$, un n-échantillon qui suis une loi de probabilité P. Soit $(x_1, ..., x_n)$, une observation de cet échantillon. La fonction de répartition F associé à P est inconnue et elle est définie par :*
$F(x) = P\{X_1 \leq x\}$ Nous estimons F par une fonction de répartition empirique, F_n. nous la définissons par :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{]-\infty;x]}(X_i)$$

Notons que $F_n : \mathbf{R} \to [0, 1]$
Alors presque sûrement (p.s), on a :

$$\sup_{x \in \mathrm{R}} |F_n(x) - F(x)| \underset{n \to +\infty}{\longrightarrow} 0$$

Nous introduisons une distance entre la fonction de répartition empirique F_n et la fonction de répartition F :

$$D_{KS} = \underset{x \in R}{Sup} \; |F_n(x) - F(x)| \xrightarrow[n \to +\infty]{p.s} 0$$

Nous avons la proposition et le théorème suivant :

Proposition 6 *Si$(X_1, ..., X_n)$ est la statistique de d'ordre associée à l'échantillon X, alors*

$$D_{KS} = \underset{i=1,...,n}{Max}\, Max\{|F(X_i) - \frac{i}{n}|, |F(X_i) - \frac{i-1}{n}|\}$$

Theorem 4 *Soit $D_n = \sqrt{n} D_{KS}$, alors D_n converge en loi vers une loi tabulée dont la fonction de répartition est donnée par :*

$$H(x) = 1 + 2\sum_{k=1}^{\infty}(-1)^k \ \exp -2k^2x^2$$

2.2.2 Application du test de Kolmornov Smirnov d'adéquation

Soit $X_i, ..., X_n$ *i.i.d* de fonction de répartition F. Nous nous donnons une fonction de répartition F^0, supposée continue. Nous testons les hypothèses suivantes au seuil $\alpha\%$:

— $H_0 : F = F^0$
— $H_1 : F \neq F^0$

Grâce au théorème (4) nous obtenons la statistique de test suivante :

$$\begin{aligned} D_n &= \sqrt{n} D_{KS} \\ &= \sqrt{n} D_{KS} = \underset{i=1,...,n}{Max}\, Max\{|F^0(X_i) - \frac{i}{n}|, |F^0(X_i) - \frac{i-1}{n}|\} \end{aligned}$$

La région de rejet associée a cette statistique de test est : $R = \{D_n > d_{n,1-\alpha}\}$ au seuil $\alpha\%$.

Pour un test unilatéral droit, les hypothèses sont :

— $H_0 : F = F^0$
— $H_1 : F < F^0$

La statistique de test est la suivante :

$$D_n^+ = \underset{x \in \mathbf{R}}{Sup}\, |(F_n(x) - F^0(x))|$$

Et une région de rejet au seuil $\alpha\%$: $R = \{D_n^+ > d_{n,1-\alpha}^+\}$

Pour un test unilatéral gauche, les hypothèses sont :

— $H_0 : F = F^0$

— $H_1 : F > F^0$

La statistique de test est la suivante :

$$D_n^- = \underset{x \in \mathbf{R}}{Sup} \; |(F_n(x) - F^0(x))|$$

Et la région de rejet au seuil $\alpha\%$: $R = \{D_n^- > d_{n,1-\alpha}^-\}$

Comme la loi D_n est tabulée sous H_0, nous trouvons les tables des quantiles $d_{n,1-\alpha}$, tels que :

$$P_{H_0}(D_n \geq d_{n,1-\alpha}) \leq \alpha$$

Nous lisons la valeur de $d_{n,1-\alpha}$ sur les tables de KS sur la figure 1 à la page 26

2.2.3 Exemple

Impact d'une campagne de publicité sur la vaccination du Covid 19

Nous avons noté pour 10 personnes vaccinés le nombre de jours écoulés entre la début de la campagne publicitaire et la date de leur vaccination. Nous obtenons les résultats suivants de l'échantillon $X = (X_1, ..., X_n)$:

i	1	2	3	4	5	6	7	8	9	10
x	0,9	28,1	1,3	7,5	48,9	2,8	15,1	9,1	12,3	66,3

La fonction de répartition F du temps écoulé entre le début de la campagne de publicité et la date de vaccination d'une personne suit une loi exponentielle de paramètre $\lambda = 0,15$. Nous cherchons à déterminer si la fonction de répartition observée de l'échantillon F_n est équivalente à la fonction de répartition théorique. Donc si F_n suit une loi exponentielle de paramètre $\lambda = 0,15$ au seuil 5%. Pour cela nous utiliser le test de Kolmorgov-Smirnov d'adéquation.

Hypothèses testées

Au seuil de 5%, nous établissons les les hypothèses suivantes :

— H_0 :F=F^0

— H_1 :F $\neq F^0$

Avec $F^0(x) = 1 - e^{-\lambda x}$

Statistiques de test

La statistique de test de Kolmorgov-Smirnov d'adéquation est la suivante :

$$D_{KS} = \sqrt{n} \underset{x \in R}{Sup} |F_n(x) - F^0(x)|$$

Où :

F_{10}, la fonction de répartition empirique associée à l'échantillon de taille 10,

F^0, la fonction de répartition de loi exponentielle de paramètre $\lambda = 0,15$

Avec $X_1, ..., X_{10}$, qui est la statistique d'ordre associée à X. nous avons la statistique de test sous H_0 suivante :

$$D_n = \underset{i=1,\ldots,10}{Max} \left\{ |F^0(X_i) - \frac{i}{10}|, |F^0(X_i) - \frac{i-1}{10}| \right\}$$

Afin de calculer le D_n nous établissons le tableau suivant :

i	x	$F^0(X_i)$	$\frac{i}{n}$	$\frac{1-i}{n}$	$Max\left\{|F^0(X_i) - \frac{i}{n}|\right\}$	$Max\left|F^0(X_i) - \frac{i-1}{n}|\right\}$
1	0,9	0,126	$\frac{1}{10} = 0,1$	$\frac{1-1}{10} = 0$	$|0,126 - 0,1| = 0,026$	$|0,126 - 0| = 0,126$
2	1,3	0,177	$\frac{2}{10} = 0,2$	$\frac{2-1}{10} = 0,1$	$|0,177 - 0,2| = 0,023$	$|0,177 - 0,1| = 0,077$
3	2,8	0,342	0,3	0,2	0,042	0,142
4	7,5	0,675	0,4	0,3	**0,275**	**0,375**
5	9,1	0,744	0,5	0,4	0,244	0,344
6	12,3	0,841	0,6	0,5	0,241	0,341
7	15,1	0,896	0,7	0,6	0,196	0,296
8	28,1	0,985	0,8	0,7	0,185	0,285
9	48,9	0,999	0,9	0,8	0,099	0,199
10	66,5	0,999	1	0,9	0,001	0,099

Nous obtenons :

$$\begin{aligned} D_{KS} &= Max(0,275; 0,375) \\ &= 0,375 \end{aligned}$$

Région de rejet

Sous H_0, D_n converge en loi vers une loi tabulée.
Nous établissons la région de rejet : $R = \{D_{KS} > d_{10;0,95}\}$

En regardant sur la table de K-S d'adéquation, nous avons : $d_{10;0,95} = 0,409$.

$0,375 < 0,409$ Donc on valide l'hypothèse nulle au risque de 5%

2.3 Test KS de comparaison de deux échantillons

2.3.1 Application du test de Kolmornov Smirnov d'homogénéité

Nous avons deux échantillons :

— Soit, $X_1, ..., X_n i.i.d$ de la fonction de répartition F_X et F_n, la fonction de répartition empirique de ce premier échantillon (X).
Avec $i = 1, ..., n$

— Soit, $Y_1, ..., Y_m i.i.d$ de la fonction de répartition F_Y et G_m, la fonction de répartition empirique de ce deuxième échantillon (Y).
Avec $j = 1, ..., m$

Les lois des variables X_i et de Y_j nous aient inconnues.

Nous testons les hypothèses suivantes au seuil $\alpha\%$:

— $H_0 : F_X = F_Y$

— $H_1 : F_X \neq F_Y$

Definition 3 La statistique de test de Kolmorgov-Smirvov d'homogénéité est défini par :

$$D_{n,m} = \sqrt{\frac{nm}{n+m}} \underset{x \in R}{Sup} |F_n(x) - G_m(x)|$$

Avec comme région de rejet au seuil $\alpha\%$: $R = \{D_{n,m} > d_{n,m,1_\alpha}\}$

Pour un test unilatéral, les hypothèses sont :

— $H_0 : F_X = F_Y$

— $H_1 : F_X \geq F_Y$

La statistique de test est la suivante :

$$D^+_{n,m} = \underset{x \in \mathbf{R}}{Sup} |(F_n(x) - G_m(x))|$$

Et la région de rejet au seuil $\alpha\%$: $R = \{D^+_{n,m} > d^+_{n,1-\alpha}\}^{\cdot}$

Proposition 7 *Si F_X est continue, la loi de $D_{n,m}$ sous l'hypothèse $F_X = F_Y$ est indépendante de F_X. Cette loi est tabulée*

Comme la loi $D_{n,m}$ est tabulée sous H_0, nous trouvons les tables des quantiles $d_{n,m,1-\alpha}$ tel que :

$$P_{H_0}(D_n \geq d_{n,1-\alpha}) \leq \alpha$$

Nous lisons la valeur de $d_{n,m,1-\alpha}$ sur les tables de KS sur la figure 2 et 3 à aux pages 27 et 28.

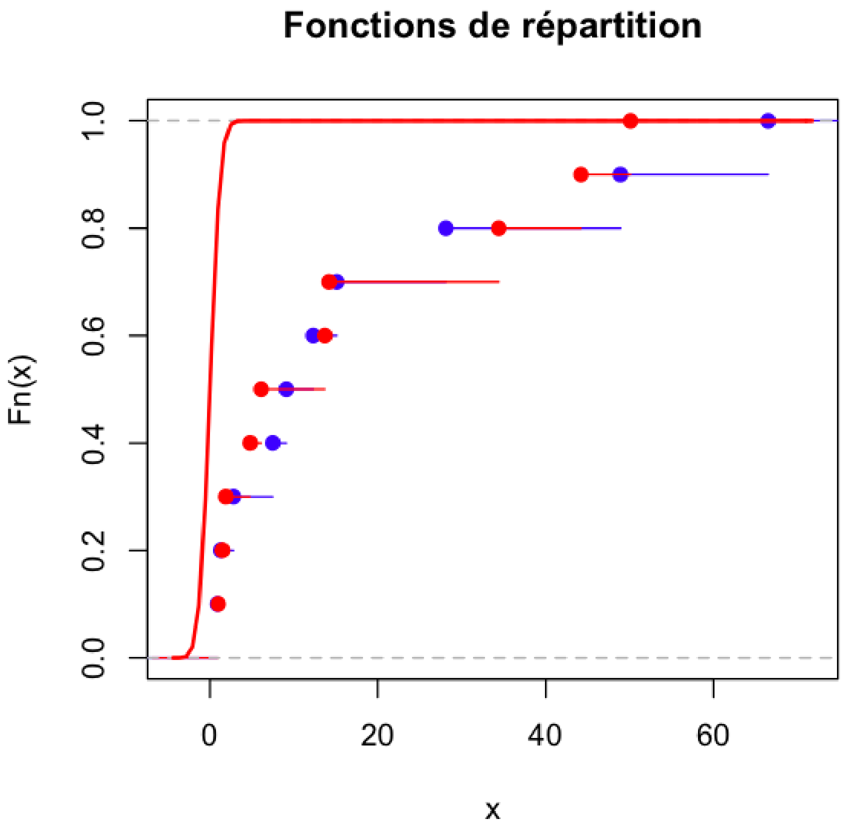

Figure 2.2 – Fonctions de répartition

2.4 Simulation du Test

KS test one sample

May 27, 2021

[1]:
```
S = c(0.9, 1.3, 2.8, 7.5, 9.1, 12.3, 15.1, 28.1, 48.9, 66.5)
length(S)
```

10

[2]:
```
for(i in 1:10){
FnF = i/10-pnorm(sort(S)[i]) # "Fn = fonction empirique", "F = fonction
 ↪théorique"
cat(FnF, ",")
}
```

```
-0.7159399 ,-0.7031995 ,-0.6974449 ,-0.6 ,-0.5 ,-0.4 ,-0.3 ,-0.2 ,-0.1 ,0 ,
```

[3]:
```
res1=c(-0.7159399 ,-0.7031995 ,-0.6974449 ,-0.6 ,-0.5 ,-0.4 ,-0.3 ,-0.2 ,-0.1 ,0)
```

[4]:
```
max(abs(res1))
```

0.7159399

[5]:
```
for(i in 1:10){
  FFn = pnorm(sort(S)[i])-(i-1)/10
  cat(FFn, ",")
}
```

```
0.8159399 ,0.8031995 ,0.7974449 ,0.7 ,0.6 ,0.5 ,0.4 ,0.3 ,0.2 ,0.1 ,
```

[6]:
```
res2=c(0.8159399 ,0.8031995 ,0.7974449 ,0.7 ,0.6 ,0.5 ,0.4 ,0.3 ,0.2 ,0.1)
max(abs(res2))
```

0.8159399

[7]:
```
D = max(max(res1,res2)) # D pour distance D
D
```

0.8159399

Critical distance (n=10, alpha=0.05) = 0.40925

D > Dc alors on rejette l'hypothèse de nullité. Les données ne suivent pas une loi normale

KS test one sample

May 27, 2021

[1]:
```
S = c(0.9, 1.3, 2.8, 7.5, 9.1, 12.3, 15.1, 28.1, 48.9, 66.5)
length(S)
```

10

[2]:
```
for(i in 1:10){
FnF = i/10-pexp(sort(S)[i], 0.15) # "Fn = fonction empirique", "F = fonction␣
↪théorique"
cat(FnF, ",")
}
```

```
-0.02628409 ,0.02283466 ,-0.04295318 ,-0.2753475 ,-0.2446193 ,-0.2419747
,-0.19617 ,-0.1852277 ,-0.0993477 ,4.654923e-05 ,
```

[3]:
```
res1=c(-0.02628409 ,0.02283466 ,-0.04295318 ,-0.2753475 ,-0.2446193 ,-0.2419747␣
↪,-0.19617 ,-0.1852277 ,-0.0993477 ,4.654923e-05)
max(abs(res1))
```

0.2753475

[4]:
```
for(i in 1:10){
  FFn = pexp(sort(S)[i], 0.15)-(i-1)/10
  cat(FFn, ",")
}
```

```
0.1262841 ,0.07716534 ,0.1429532 ,0.3753475 ,0.3446193 ,0.3419747 ,0.29617
,0.2852277 ,0.1993477 ,0.09995345 ,
```

[5]:
```
res2=c(0.1262841 ,0.07716534 ,0.1429532 ,0.3753475 ,0.3446193 ,0.3419747 ,0.
↪29617 ,0.2852277 ,0.1993477 ,0.09995345)
max(abs(res2))
```

0.3753475

[7]:
```
D = max(max(res1,res2)) # D pour distance D
D
```

0.3753475

Critical distance (n=10, alpha=0.05) = 0.40925

D < Dc alors on accepte l'hypothèse de nullité. Les données suivent une loi exponentielle de paramètre lambda = 0.15

0.1 Test avec la fonction ks.test

[8]:
```
KS = ks.test(S, "pexp", 0.15)
KS
```

```
        One-sample Kolmogorov-Smirnov test

data:  S
D = 0.37535, p-value = 0.08971
alternative hypothesis: two-sided
```

D < Dc alors on accepte l'hypothèse de nullité. Les données suivent une loi exponentielle de paramètre lambda = 0.15

Chapitre 3

Chapitre 3 : Application à l'actualité

Malgré nos longues recherches, nous n'avons pas trouvé de données suffisantes liées à l'impact du Covid sur le secteur culturel. Pour une application optimale et un résultat fiable nous avons donc utilisé une base de données qui traite le prix des appartements à Taipei en Taïwan avec 301 variables. Le code sera rédigé avec Python et R sur jupyter notebook et publié dans un fichier à part.

Annexe

.1 KS test

La table nous donne la distribution de l'échantillonnage de D_n par $P(D_n \leq d_{n,1-\alpha}) = \alpha$

One-Sided Test $1-\alpha =$	**0.90**	**0.95**	**0.975**	**0.99**	**0.995**	$1-\alpha =$	**0.90**	**0.95**	**0.975**	**0.99**	**0.995**
Two-Sided Test $1-\alpha =$	**0.80**	**0.90**	**0.95**	**0.98**	**0.99**	$1-\alpha =$	**0.80**	**0.90**	**0.95**	**0.98**	**0.99**
$n = 1$	0.900	0.950	0.975	0.990	0.995	$n = 21$	0.226	0.259	0.287	0.321	0.344
2	0.684	0.776	0.842	0.900	0.929	**22**	0.221	0.253	0.281	0.314	0.337
3	0.565	0.636	0.708	0.785	0.829	**23**	0.216	0.247	0.275	0.307	0.330
4	0.493	0.565	0.624	0.689	0.734	**24**	0.212	0.242	0.269	0.301	0.323
5	0.447	0.509	0.563	0.627	0.669	**25**	0.208	0.238	0.264	0.295	0.317
6	0.410	0.468	0.519	0.577	0.617	**26**	0.204	0.233	0.259	0.290	0.311
7	0.381	0.436	0.483	0.538	0.576	**27**	0.200	0.229	0.254	0.284	0.305
8	0.358	0.410	0.454	0.507	0.542	**28**	0.197	0.225	0.250	0.279	0.300
9	0.339	0.387	0.430	0.480	0.513	**29**	0.193	0.221	0.246	0.275	0.295
10	0.323	0.369	0.409	0.457	0.489	**30**	0.190	0.218	0.242	0.270	0.290
11	0.308	0.352	0.391	0.437	0.468	**31**	0.187	0.214	0.238	0.266	0.285
12	0.296	0.338	0.375	0.419	0.449	**32**	0.184	0.211	0.234	0.262	0.281
13	0.285	0.325	0.361	0.404	0.432	**33**	0.182	0.208	0.231	0.258	0.277
14	0.275	0.314	0.349	0.390	0.418	**34**	0.179	0.205	0.227	0.254	0.273
15	0.266	0.304	0.338	0.377	0.404	**35**	0.177	0.202	0.224	0.251	0.269
16	0.258	0.295	0.327	0.366	0.392	**36**	0.174	0.199	0.221	0.247	0.265
17	0.250	0.286	0.318	0.355	0.381	**37**	0.172	0.196	0.218	0.244	0.262
18	0.244	0.279	0.309	0.346	0.371	**38**	0.170	0.194	0.215	0.241	0.258
19	0.237	0.271	0.301	0.337	0.361	**39**	0.168	0.191	0.213	0.238	0.255
20	0.232	0.265	0.294	0.329	0.352	**40**	0.165	0.189	0.210	0.235	0.252

FIGURE 1 – Table des quantiles de la statistique de test de KS

La table nous donne la distribution de l'échantillonnage de $D_{n,m}$ par

$$P(D_{n,m} \leq d_{n,m,1-\alpha}) = \alpha$$

One-Sided Test $1-\alpha =$	**0.90**	**0.95**	**0.975**	**0.99**	**0.995**	$1-\alpha =$	**0.90**	**0.95**	**0.975**	**0.99**	**0.995**
Two-Sided Test $1-\alpha =$	**0.80**	**0.90**	**0.95**	**0.98**	**0.99**	$1-\alpha =$	**0.80**	**0.90**	**0.95**	**0.98**	**0.99**
$n = 3$	2/3	2/3				$n = 20$	6/20	7/20	8/20	9/20	10/20
4	3/4	3/4	3/4			**21**	6/21	7/21	8/21	9/21	10/21
5	3/5	3/5	4/5	4/5	4/5	**22**	7/22	8/22	8/22	10/22	10/22
6	3/6	4/6	4/6	5/6	5/6	**23**	7/23	8/23	9/23	10/23	10/23
7	4/7	4/7	5/7	5/7	5/7	**24**	7/24	8/24	9/24	10/24	11/24
8	4/8	4/8	5/8	5/8	6/8	**25**	7/25	8/25	9/25	10/25	11/25
9	4/9	5/9	5/9	6/9	6/9	**26**	7/26	8/26	9/26	10/26	11/26
10	4/10	5/10	6/10	6/10	7/10	**27**	7/27	8/27	9/27	11/27	11/27
11	5/11	5/11	6/11	7/11	7/11	**28**	8/28	9/28	10/28	11/28	12/28
12	5/12	5/12	6/12	7/12	7/12	**29**	8/29	9/29	10/29	11/29	12/29
13	5/13	6/13	6/13	7/13	8/13	**30**	8/30	9/30	10/30	11/30	12/30
14	5/14	6/14	7/14	7/14	8/14	**31**	8/31	9/31	10/31	11/31	12/31
15	5/15	6/15	7/15	8/15	8/15	**32**	8/32	9/32	10/32	12/32	12/32
16	6/16	6/16	6/25	8/16	12/15	**34**	8/34	10/34	11/34	12/34	13/34
17	9/29	7/17	7/17	8/22	9/17	**36**	9/36	10/36	11/36	12/36	13/36
18	6/18	7/18	8/18	9/18	9/19	**38**	9/38	10/38	11/38	13/38	14/38
19	6/19	7/19	8/19	9/19	9/19	**40**	9/40	10/40	12/40	13/40	14/40
						Approximation for $n > 40$:	$\frac{1.52}{\sqrt{n}}$	$\frac{1.73}{\sqrt{n}}$	$\frac{1.92}{\sqrt{n}}$	$\frac{2.15}{\sqrt{n}}$	$\frac{2.30}{\sqrt{n}}$

FIGURE 2 – Table des quantiles de la statistique de test de KS quand n=m

[13] [14] [12] [19] [17] [18] [16] [20] [15]
[10] [9] [3] [1] [6] [5] [4] [11] [7] [8] [2]

La table nous donne la distribution de l'échantillonnage de $D_{n,m}$ par
$P(D_{n,m} \leq d_{n,m,1-\alpha}) = \alpha$

One-Sided Test	**$1-\alpha=$**	**0.90**	**0.95**	**0.975**	**0.99**	**0.995**
Two-Sided Test	**$1-\alpha=$**	**0.80**	**0.90**	**0.950**	**0.98**	**0.990**
$n=1$	**$m=9$**	17/18				
	10	9/10				
$n=2$	**3**	5/6				
	4	3/4				
	5	4/5	4/5			
	6	5/6	5/6			
	7	5/7	6/7			
	8	3/4	7/8	7/8		
	9	7/9	8/9	8/9		
	10	7/10	4/5	9/10		
$n=3$	**$m=4$**	3/4	3/4			
	5	2/3	4/5	4/5		
	6	2/3	2/3	5/6		
	7	2/3	5/7	6/7	6/7	
	8	5/8	3/4	3/4	7/8	
	9	2/3	2/3	7/9	8/9	8/9
	10	3/5	7/10	4/5	9/10	9/10
	12	7/12	2/3	3/4	5/6	11/12
$n=4$	**$m=5$**	3/5	3/4	4/5	4/5	
	6	7/12	2/3	3/4	5/6	5/6
	7	17/28	5/7	3/4	6/7	6/7
	8	5/8	5/8	3/4	7/8	7/8
	9	5/9	2/3	3/4	7/9	8/9
	10	11/20	13/20	7/10	4/5	4/5
	12	7/12	2/3	2/3	3/4	5/6
	16	9/16	5/8	11/16	3/4	13/16
$n=5$	**$m=6$**	3/5	2/3	2/3	5/6	5/6
	7	4/7	23/35	5/7	29/35	6/7
	8	11/20	5/8	27/40	4/5	4/5
	9	5/9	3/5	31/45	7/9	4/5
	10	1/2	3/5	7/10	7/10	4/5
	15	8/15	3/5	2/3	11/15	11/15
	20	1/2	11/20	3/5	7/10	3/4

One-Sided Test	**$1-\alpha=$**	**0.90**	**0.95**	**0.975**	**0.99**	**0.995**
Two-Sided Test	**$1-\alpha=$**	**0.80**	**0.90**	**0.950**	**0.98**	**0.990**
$n=6$	**$m=7$**	23/42	4/7	29/42	5/7	5/6
	8	1/2	7/12	2/3	3/4	3/4
	9	1/2	5/9	2/3	13/18	7/9
	10	1/2	17/30	19/30	7/10	11/15
	12	1/2	7/12	7/12	2/3	3/4
	18	4/9	5/9	11/18	2/3	13/18
	24	11/24	1/2	7/12	5/8	2/3
$n=7$	**$m=8$**	27/56	33/56	5/8	41/56	3/4
	9	31/63	5/9	40/63	5/7	47/63
	10	33/70	39/70	43/70	7/10	5/7
	14	3/7	1/2	4/7	9/14	5/7
	28	3/7	13/28	15/28	17/28	9/14
$n=8$	**$m=9$**	4/9	13/24	5/8	2/3	3/4
	10	19/40	21/40	23/40	27/40	7/10
	12	11/24	1/2	7/12	5/8	2/3
	16	7/16	1/2	9/16	5/8	5/8
	32	13/32	7/16	1/2	9/16	19/32
$n=9$	**$m=10$**	7/15	1/2	26/45	2/3	31/45
	12	4/9	1/2	5/9	11/18	2/3
	15	19/45	22/45	8/15	3/5	29/45
	18	7/18	4/9	1/2	5/9	11/18
	36	13/36	5/12	17/36	19/36	5/9
$n=10$	**$m=15$**	2/5	7/15	1/2	17/30	19/30
	20	2/5	9/20	1/2	11/20	3/5
	40	7/20	2/5	9/20	1/2	
$n=12$	**$m=15$**	23/60	9/20	1/2	11/20	7/12
	16	3/8	7/16	23/48	13/24	7/12
	18	13/36	5/12	17/36	19/36	5/9
	20	11/30	5/12	7/15	31/60	17/30
$n=15$	**$m=20$**	7/20	2/5	13/30	29/60	31/60
$n=16$	**$m=20$**	27/80	31/80	17/40	19/40	41/80
Large-sample approximation		$1.07\sqrt{\frac{m+n}{mn}}$	$1.22\sqrt{\frac{m+n}{mn}}$	$1.36\sqrt{\frac{m+n}{mn}}$	$1.52\sqrt{\frac{m+n}{mn}}$	$1.63\sqrt{\frac{m+n}{mn}}$

FIGURE 3 – Table des quantiles de la statistique de test de KS quand $n \neq m$

Bibliographie

[1] *https ://docplayer.fr/15141400-Tests-statistiques-rejeter-ne-pas-rejeter-se-risquer-magalie-fromont-annee-universitaire-2015-2016.html.*

[2] *https ://github.com/kimfetti/Videos/blob/master/Seaborn/02*$_K$*DEplot.ipynb.*

[3] *https ://github.com/statsmodels/statsmodels/blob/main/statsmodels/nonparametric/bandwidths.py.*

[4] *https ://gsalvatovallverdu.gitlab.io/python/kernel*$_d$*ensity*$_e$*stimation/.*

[5] *https ://jakevdp.github.io/PythonDataScienceHandbook/05.13-kernel-density-estimation.html.*

[6] *https ://kdepy.readthedocs.io/en/latest/introduction.html.*

[7] *https ://scikit-learn.org/stable/auto*$_e$*xamples/neighbors/plot*$_k$*de*$_1$*d.html.*

[8] *https ://seaborn.pydata.org/generated/seaborn.kdeplot.html.*

[9] *https ://sparky.rice.edu//astr360/kstest.pdf.*

[10] *https ://www.rdocumentation.org/packages/dgof/versions/1.2/topics/ks.test.*

[11] *https ://www.statsmodels.org/stable/examples/notebooks/generated/kernel*$_d$*ensity.html.*

[12] STAT 2413. *Chapitre 3 Estimation non-paramétrique d'une fonction de répartition et d'une densité*. 2002-2003.

[13] Arnak S. DALALYAN. *STATISTIQUE AVANCÉE : MÉTHODES NON-PARAMÉTRIQUES.*

[14] Lamia Ferhat. *Estimation d'une fonction de densité par la méthode des noyaux et application à la VaR.* PhD thesis, UMMTO, 2012.

[15] Christian Gagné.

[16] https ://docplayer.fr/15141400-Tests-statistiques-rejeter-ne-pas-rejeter-se-risquer-magalie-fromont-annee-universitaire 2015-2016.html. *Tests Statistiques.*

[17] https ://lemakistatheux.wordpress.com/2013/05/09/le-test-de-kolmogorov smirnov/. *Le test de Kolmogorov-Smirnov.*

[18] https ://www.math.univ-toulouse.fr/ besse/Wikistat/pdf/st-m-inf np.pdf. *Tests non paramétriques.*

[19] Statistique Mathématique. *TD 8 : tests de Kolmogorov-Smirnov.* 2019-2020.

[20] Alexandre B. Tsybakov. *Introduction à l'estimation non-paramétrique.*