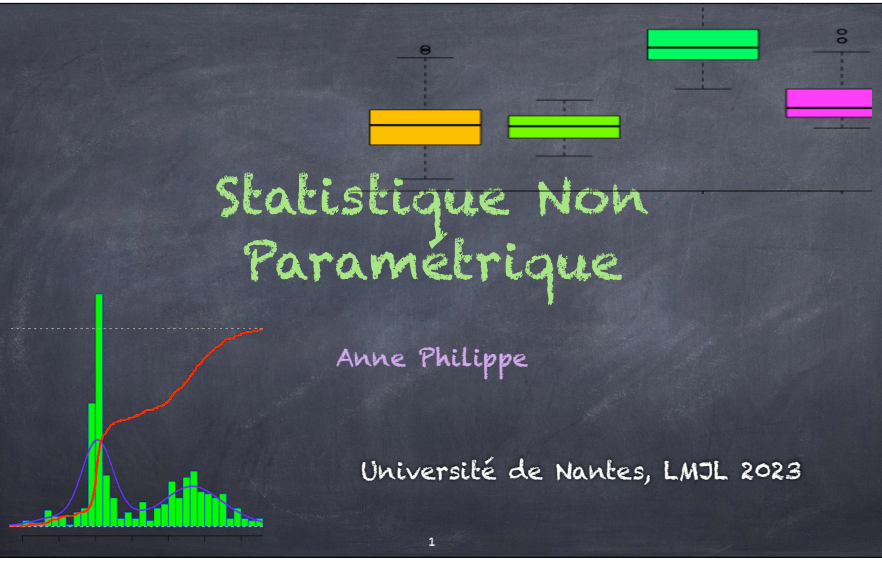


Statistique Non Paramétrique

Anne Philippe

Université de Nantes, LMJL 2023

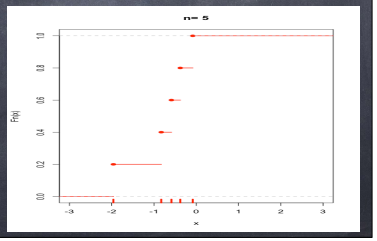
- I - processus empirique

Hypothèse (H) : Soit $(X_1, \dots X_n)$ un n-échantillon de variables aléatoires iid suivant F (fonction de répartition de la loi commune)

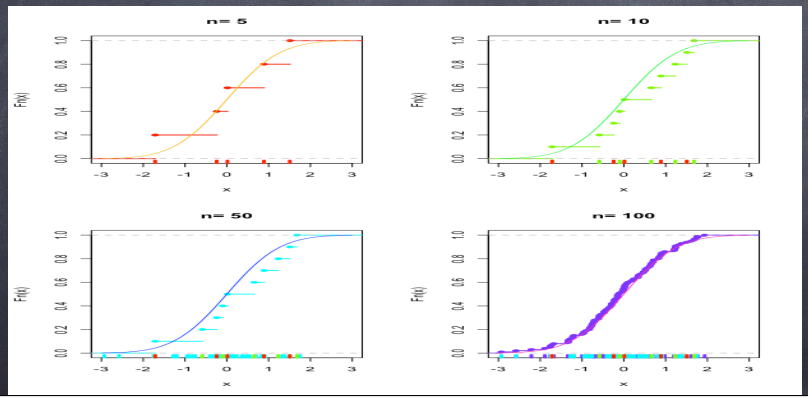
- **Définition** : Le processus empirique est une fonction aléatoire définie par

$$F_n(t,\omega) = F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{]-\infty,t]}(X_i(\omega))$$

  - $F_n(t)$ représente la proportion des observations inférieures à t
  - $F_n(t) = \begin{cases} 0 \text{ si } t < \min(X_1, \dots X_n) \\ 1 \text{ si } t \geq \max(X_1, \dots X_n) \end{cases}$
  - C'est une fonction constante par morceaux



- **Théorème de Glivenko Cantelli** : Sous l'hypothèse (H) on a $\sup_{t\in R} |F_n(t) - F(t)| \xrightarrow[ps]{n\to\infty} 0$

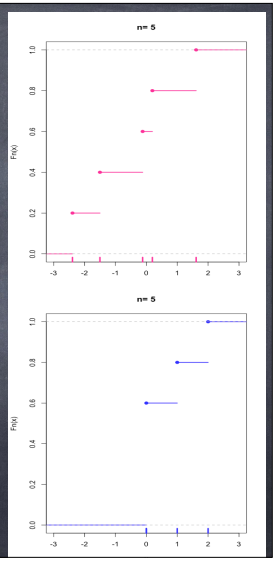


- **Conséquence du Th de G-C** :
  $\forall t \in R, F_n(t)$ est un estimateur fortement consistant de $F(t)$
- **Conséquence du TCL** : $\sqrt{n}(F_n(t) - F(t)) \xrightarrow[loi]{n\to\infty} \mathcal{N}(0,\sigma_t^2)$
  donc $\forall t \in R, F_n(t)$ est un estimateur $\sqrt{n}-$consistant de $F(t)$
- **Conséquence du Th de Slutsky** : Si $F(t) \neq 0$ et $F(t) \neq 1$ alors
  $$\frac{\sqrt{n}(F_n(t) - F(t))}{\sqrt{F_n(t)(1-F_n(t))}} \xrightarrow[loi]{n\to\infty} \mathcal{N}(0,1)$$

Le processus empirique F_n est la fonction de répartition d'une loi discrète de support les valeurs prises par $\{X_1(\omega),\dots X_n(\omega)\}$

I. Si $\{X_1(\omega),\dots X_n(\omega)\}$ sont tous distincts,
F_n est la fct de répartition de la loi uniforme sur $\{X_1(\omega),\dots X_n(\omega)\}$

II. Sinon on note $\{\tilde{X}_1(\omega),\dots,\tilde{X}_m(\omega)\}$ l'échantillon sans répétition et $\hat{p}_j = \frac{1}{n} card\{i : X_i = \tilde{X}_j\} \quad j-1,\dots,m.$
F_n est la fct de répartition de la loi discrète de support $\{\tilde{X}_1(\omega),\dots,\tilde{X}_m(\omega)\}$ et de proba $\{\hat{p}_1,\dots,\hat{p}_m\}$



Moments de F_n

- F_n définit une loi discrète finie donc cette loi admet des moments de tout ordre
- Le moment d'ordre k de F_n est égal à $\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$
  - Pour k=1 on retrouve la moyenne empirique $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
  - $\hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ est la variance empirique

Application : estimation par injection

Soit h une application telle que $h(X_1) \in L^1$
On veut estimer $\theta_h(F) = E(h(X_1))$.

Définition L'estimateur par injection du paramètre $\theta_h(F)$ est $\theta_h(F_n)$. Il est égal à $\theta_h(F_n) = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$

Propriétés Sous l'hypothèse H

1. Si $h(X_1) \in L^1$, $\theta_h(F_n)$ est un estimateur sans biais fortement consistant de $\theta_h(F)$

2. Si $h(X_1) \in L^2$ alors c'est aussi un estimateur consistant au sens L^2 et il est $\sqrt{n}-$consistant.

Exemple : estimation des moments

On suppose que $X_1 \in L^k$. On note μ_k le moment d'ordre k
L'estimateur par injection de μ_k est $\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$

Propriétés Sous l'hypothèse H

1. Si $X_1 \in L^p$, $\hat{\mu} = (\hat{\mu}_1,\dots,\hat{\mu}_p)$ est un estimateur sans biais fortement consistant de $\mu = (\mu_1,\dots,\mu_p)$

2. Si $X_1 \in L^{2p}$ alors c'est aussi un estimateur consistant au sens L^2 et il est $\sqrt{n}-$consistant
on a $\sqrt{n}(\hat{\mu}-\mu) \xrightarrow{loi} \mathcal{N}_p(0,\Sigma_\mu)$ où Σ_μ est définie par $\Sigma_{i,j} = \mu_{i+j} - \mu_i\mu_j$

Méthode des moments

- On suppose que H est vérifiée et la loi commune appartient à la famille paramétrique $\mathcal{F} = \{P_\theta, \theta \in \Theta\}$.
- On veut estimer le paramètre θ
- On suppose que $X_1 \in L^P$ et qu'il existe $g : R^p \to \Theta$ tel que $\theta = g(\mu_1, \ldots, \mu_p)$
- Définition L'estimateur des moments du paramètre θ est défini par $\hat{\theta}_n^M = g(\hat{\mu}_1, \ldots, \hat{\mu}_p)$
- Exemple : modèle exponentiel $g(x) = 1/x$

Méthode des moments (cont.)

Sous l'hypothèse H et $P_{X_1} \in \{P_\theta, \theta \in \Theta\}$

1. Si $X_1 \in L^p$ et g est continue sur R^p alors $\hat{\theta}_n^M$ est un estimateur fortement consistant de θ
2. Si $X_1 \in L^{2p}$ et g est C^1 sur R^p alors $\hat{\theta}_n^M$ est $\sqrt{n}-$consistant on a $\sqrt{n}(\hat{\theta}_n^M - \theta) \xrightarrow{loi} \mathcal{N}_k(0, \Lambda)$ avec $\Lambda = Dg(\mu)\Sigma_\mu Dg(\mu)^T$

Remarques sur la méthode des moments

Il n'y a pas unicité de l'estimateur des moments

Choix de p ? Choix de g ?

Dimension 1 : On cherchera l'estimateur des moments qui possède la plus petite variance asymptotique Λ.

Variance de la loi limite $\Lambda = Dg(\mu)\Sigma_\mu Dg(\mu)^T$

1. L'estimateur des moments de la matrice de variance-covariance Σ_μ est donné par $\hat{\Sigma}_\mu$ avec $\hat{\Sigma}_{i,j} = \hat{\mu}_{i+j} - \hat{\mu}_i\hat{\mu}_j$.
   Il est consistant sous l'hypothèse $X_1 \in L^{2p}$
2. $Dg(\hat{\mu})$ est un estimateur consistant de $Dg(\mu)$ sous l'hypothèse car la fonction g est C^1
3. On obtient un estimateur consistant de Λ en prenant $\hat{\Lambda} = Dg(\hat{\mu})\hat{\Sigma}_\mu Dg(\hat{\mu})^T$
4. En dimension 1 : on a $\dfrac{\sqrt{n}}{\sqrt{\hat{\Lambda}}}(\hat{\theta}_n^M - \theta) \xrightarrow{loi} \mathcal{N}(0,1)$

   [application Intervalle de confiance asymptotique] 

II - processus quantile

- Si F est inversible, la fonction quantile est $Q(u) = F^{-1}(u)$

  Sinon c'est le pseudo inverse $Q(u) = F^{-}(u) = inf\{x \ : \ F(x) \geq u\}$

  $Q(u)$ est le quantile d'ordre u (u=1/2 -> médiane)

- Le processus empirique n'est jamais une fonction inversible mais on peut calculer son pseudo inverse
- Définition : L'estimateur par injection du quantile d'ordre u est $Q_n(u) = F_n^{-}(u)$
- Q_n est le processus quantile empirique

Convergence de Q_n

- Si F est continue et inversible alors $Q_n(u) \xrightarrow[ps]{n\to\infty} F^{-1}(u)$
- (Admis) Si F est dérivable au point $F^{-1}(u)$ avec une dérivée f strictement positive alors

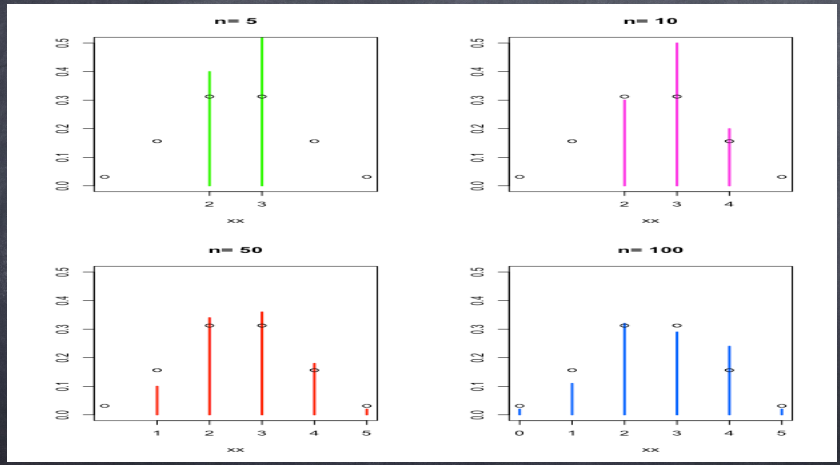
$$\sqrt{n}(F_n^{-}(u) - F^{-1}(u)) \xrightarrow{loi} \mathcal{N}\left(0, \frac{u(1-u)}{f^2(F^{-1}(u))}\right)$$

- Remarque
  - L'estimation de la variance nécessite une estimation consistante de la densité f.

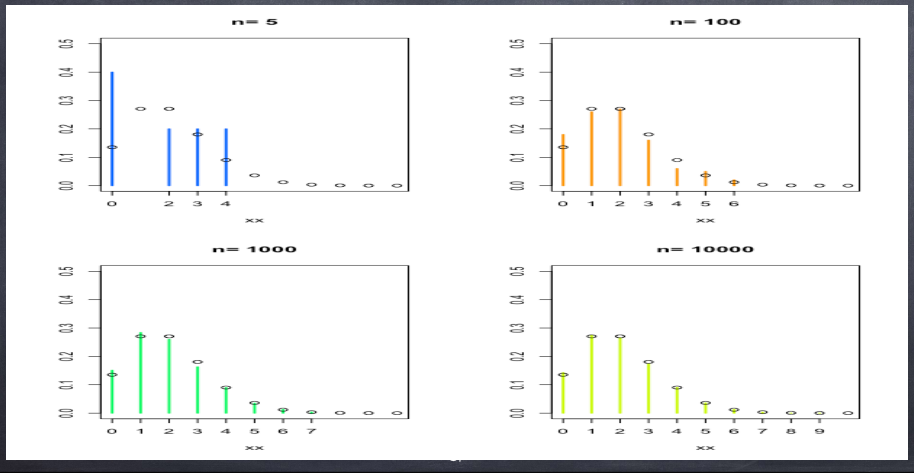
III- estimation d'une loi discrète

- On suppose que H est vérifiée et la loi de X_1 est discrète à valeurs dans E.
- On veut estimer pour tout $e \in E : P(X_1 = e) = p(e)$
- Le processus empirique définit une loi discrète de support $\{\tilde{X}_1(\omega), \ldots, \tilde{X}_m(\omega)\}$ et de probabilités $\hat{p}_j = \frac{1}{n} card\{i : X_i = \tilde{X}_j\}$.
- Pour tout $e \in E$, il existe $n \in N$: $\forall m \geq n$ et $e \in \{\tilde{X}_1(\omega), \ldots, \tilde{X}_m(\omega)\}$.
- Proposition : La loi discrète du processus empirique est un estimateur sans biais , fortement consistant , consistant au sens L^2 et $\sqrt{n}-$ consistant de la loi de commune

Loi binomiale n = 5 et p=1/2



Loi de Poisson de paramètre 2



IV- Estimation d'une densité

- On suppose que H est vérifiée et la loi de X_1 admet une densité f
- On veut estimer la fonction f
- Remarque : Comme le processus empirique n'admet pas de densité, on ne peut pas en déduire un estimateur de la densité de X_1

Histogramme

- Rappel : Si F est dérivable par morceaux on a
$$f(x) = \lim_{h\to 0} \frac{1}{2h}(F(x+h) - F(x-h))$$
- A (h, x) fixé : $\frac{1}{2h}(F_n(x+h) - F_n(x-h))$ est un estimateur de $\frac{1}{2h}(F(x+h) - F(x-h))$
Il est sans biais et fortement consistant

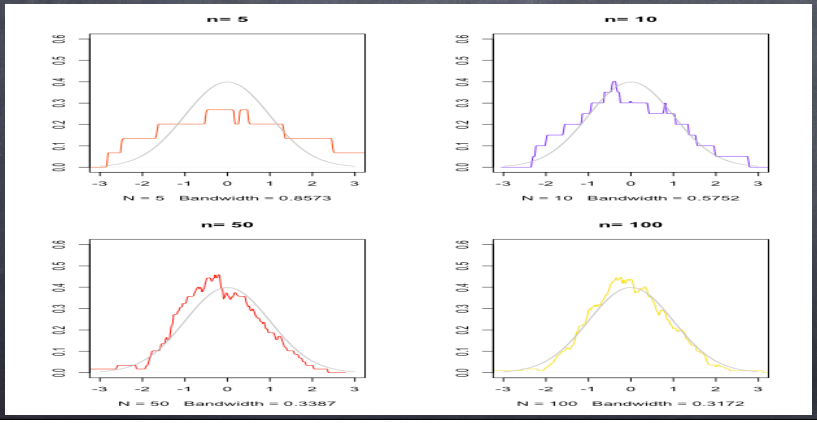
Propriétés de l'histogramme

- Définition : on définit l'histogramme mobile par
$$f_n^H(x) = \frac{1}{2h}(F_n(x+h) - F_n(x-h)) = \frac{1}{2nh}\sum_{i=1}^{n} 1_{]x-h,x+h]}(X_i)$$
où h est la fenêtre de l'histogramme.
- On choisit une fenêtre qui dépend du nombre d'observation n
Soit h_n une suite réelles positive
Théorème :

1. Si $h = h_n \xrightarrow{n\to\infty} 0$ alors $f_n^H(x)$ est un estimateur asymptotiquement sans biais de la densité $f(x)$
2. Si $h = h_n \xrightarrow{n\to\infty} 0$ et $nh_n \xrightarrow{n\to\infty} \infty$ alors $f_n^H(x)$ est un estimateur L^2 consistant de la densité $f(x)$

Pour différentes valeurs de n :

on représente l'histogramme d'un n-échantillon de variables iid suint la loi gaussienne standard N(0,1)



Estimateur à noyau de la densité

- L'idée est de régulariser l'estimateur de la densité pour obtenir un estimateur continu.
- Remarque : on a

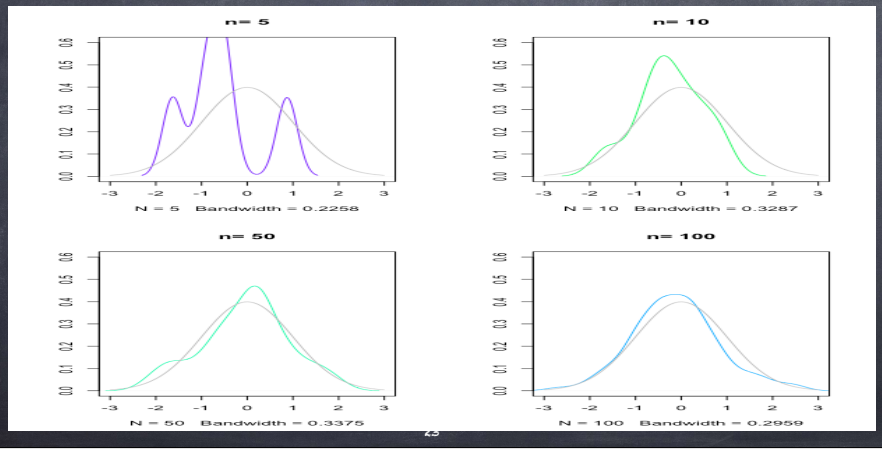
$$f_n^H(x) = \frac{1}{2nh_n}\sum_{i=1}^{n} 1_{]-1,1]}\left(\frac{x-X_i}{h_n}\right) = \frac{1}{nh_n}\sum_{i=1}^{n} K_U\left(\frac{x-X_i}{h_n}\right)$$

où K_U est la densité de la loi uniforme sur [-1,1]

- On remplace K_U par une densité K telle que K est paire et les fonctions K^2 et $x \to x^2K(x)$ sont intégrables
- Exemple de noyau : la densité de la loi gaussienne standard

On calcule sur les n-échantillon précédent

L'estimateur à noyau avec le noyau gaussien



Estimateur à noyau

- L'estimateur à noyau de la densité f est défini par

$$f_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)$$

K : le noyau est une densité paire et les fonctions K^2 et $x \to x^2K(x)$ sont intégrables

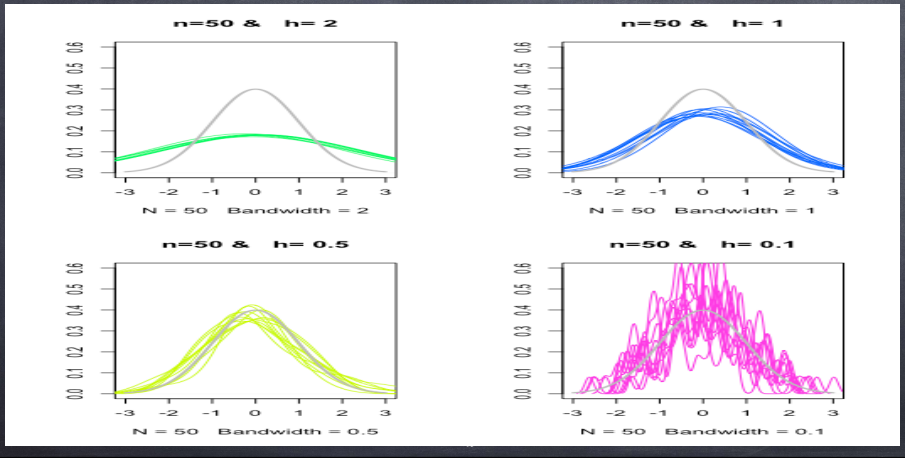
h_n est une suite de réels positifs appelée fenêtre.

Théorème : On suppose que la densité f est C^2 et f, f', f'' sont bornées, strictement positives

1. Si $h = h_n \xrightarrow{n\to\infty} 0$ alors $f_n(x)$ est un estimateur asymptotiquement sans biais de la densité $f(x)$

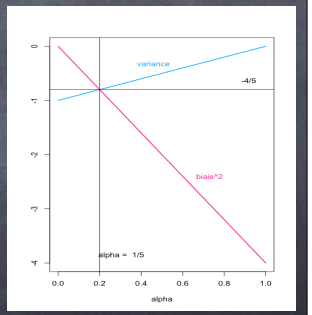
2. Si $h = h_n \xrightarrow{n\to\infty} 0$ et $nh_n \xrightarrow{n\to\infty} \infty$ alors $f_n(x)$ est un estimateur L^2 consistant de la densité $f(x)$

Effet de h (10 réplications)



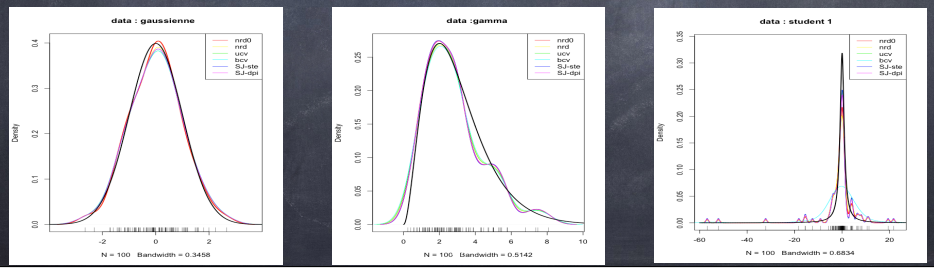
Compromis biais – variance

- Le biais est équivalent à Ch_n^2
- La variance est équivalente à $C'\frac{1}{nh_n}$
- L'erreur L^2 est équivalente à $C^2h_n^4 + C'\frac{1}{nh_n}$
- Proposition Si $h_n = cn^{-\alpha}$ alors le choix optimal au sens de l'erreur quadratique est $\alpha = \frac{1}{5}$. L'erreur L^2 est équivalente à $Cn^{-4/5}$



choix de la constante
Critère de Silverman

- Les constantes sont explicites mais elles dépendent de la densité qui est inconnue
- Critère de Silverman : on calcule les constantes sous l'hypothèses que la densité est celle d'une loi gaussienne.



A partir de l'estimateur à noyau de la densité f_n on peut estimer

- La fonction de répartition en prenant $F_n^{noyau}(x) = \int_{-\infty}^{x} f_n(t)\,dt$
- $E(h(X_1))$ en prenant $\int h(t)f_n(t)\,dt$ si $h(X_1) \in L^1$ et $h(Y) \in L^1$ avec $Y \sim K$.
  - Cas particulier la moyenne, la variance etc
- La fonction quantile en prenant l'inverse ou le pseudo inverse de F_n^{noyau}