Partiel de Statistique

Mercredi 16 janvier 2008, 8h00-10h00

INP ENSEEIHT

Partiel sans documents (une feuille manuscrite A4 est autorisée)

RAPPELS LOI GAMMA

LOI	Densité de probabilité	moyenne	variance	fonction caractéristique.
Gamma $\Gamma(\theta,\nu)$	$f(x)=\frac{\theta^\nu}{\Gamma(\nu)}e^{-\theta x}x^{\nu-1}$ $\theta>0,\ \nu>0$ $x\geq 0$	$\frac{\nu}{\theta}$	$\frac{\nu}{\theta^2}$	$\frac{1}{\left(1-i\frac{t}{\theta}\right)^\nu}$

RAPPELS FONCTION GAMMA

$$\Gamma(n)=\int_0^\infty u^{n-1}e^{-u}du=(n-1)!$$

Exercice 1 : Estimation

On considère une variable aléatoire X de loi de Pareto (ce que l'on notera dans la suite $X\sim P(a)$) de densité

$$h(x)=\frac{a}{x^{a+1}}\mathbb{I}_{[1,\infty[}(x),$$

où $a>0$ est un paramètre inconnu et où $\mathbb{I}_{[1,\infty[}(x)$ est la fonction indicatrice sur le domaine $[1,\infty[$ ($\mathbb{I}_{[1,\infty[}(x)=1$ si $x\in[1,\infty[$ et $\mathbb{I}_D(x)=0$ sinon). Dans cette partie, on dispose de n observations $x_1,...,x_n$ associées à cette loi de Pareto et on cherche à estimer le paramètre a.

1) Montrer que l'estimateur du maximum de vraisemblance de a s'écrit (en n'oubliant pas de faire un tableau de variation)

$$\widehat{a}_{MV}=\frac{n}{\sum_{i=1}^n \ln X_i}.$$

- Montrer que la variable aléatoire $Z=\ln X$ suit une loi gamma dont on déterminera les paramètres.
- Déterminer la fonction caractéristique de $U=\sum_{i=1}^n Z_i=\sum_{i=1}^n \ln X_i$ lorsque $(X_1,...,X_n)$ est un échantillon de loi de Pareto $P(a)$. En déduire que U suit une loi gamma de paramètres a et n, i.e. $U\sim\Gamma(a,n)$. Déterminer la moyenne et la variance de la variable aléatoire $\frac{1}{U}$ notées $E\left[\frac{1}{U}\right]$ et $V\left[\frac{1}{U}\right]$.
- Déterminer le biais de l'estimateur $\widehat{a}_{MV}$. En déduire un estimateur non biaisé de a noté $\widetilde{a}$ et déterminer la variance de cet estimateur.
- L'estimateur $\widetilde{a}$ est-il l'estimateur efficace de a ?

2) On suppose maintenant qu'on dispose d'une information a priori sur le paramètre a résumée dans la densité a priori

$$g(a)=\lambda e^{-\lambda a}\mathbb{I}_{[0,\infty[}(a).$$

- Montrer que la densité a posteriori du paramètre a notée $f(a|x_1,...,x_n)$ est la densité d'une loi gamma dont on déterminera les paramètres.
- Déterminer l'estimateur MAP du paramètre a et étudier son comportement lorsque n est "grand".
- Déterminer l'estimateur MMSE du paramètre a.

Exercice 2 : Tests Statistiques

On dispose toujours de n observations $x_1, ..., x_n$ associées à une loi de Pareto $P(a)$ et on considère le test d'hypothèses simples

$$\begin{aligned} H_0 &: a = a_0 \\ H_1 &: a = a_1 \text{ avec } a_1 > a_0 > 0 \end{aligned}$$

1) Déterminer la statistique du test de Neyman-Pearson $U(X_1, ..., X_n)$ (notée U pour simplifier) et la zone de rejet de H_0 issue de ce test. On suppose dans ce qui suit que U suit une loi gamma $\Gamma(a_0, n)$ sous l'hypothèse H_0 et une loi gamma $\Gamma(a_1, n)$ sous l'hypothèse H_1.

- Montrer que le seuil du test de Neyman-Pearson noté S_α s'exprime en fonction de α sous la forme suivante

$$S_\alpha = \frac{1}{a_0} I_n^{-1}(\alpha) \text{ avec } I_n(x) = \frac{1}{\Gamma(n)} \int_0^x e^{-u} u^{n-1} du$$

- Exprimer la puissance du test notée π en fonction de a_1, S_α et $I_n(x)$.
- Déterminer les courbes caractéristiques opérationnelles du récepteur (courbes COR) pour le test considéré ci-dessus et en déduire comment la performance du test dépend des paramètres a_0 et a_1. Représenter la forme des courbes COR pour $(a_0, a_1) = (1, 2)$ et pour $(a_0, a_1) = (1, 3)$. Commentaires.

2) On désire faire un test de Kolmogorov pour tester si les observations $x_1, ..., x_n$ sont issues d'une loi de Pareto de paramètre $a_0 = 1$, c'est-à-dire

$$\begin{aligned} H_0 &: X_i \sim P(1) \\ H_1 &: \text{non } H_0 \end{aligned}$$

- Déterminer et tracer la fonction de répartition d'une loi de Pareto $P(1)$ notée $F_0(x)$.
- Tracer la fonction de répartition $\widehat{F}(x)$ des observations (pour simplifier les calculs, les observations sont entières mais ce serait très peu probable en pratique !)

$$x_1 = 2, x_2 = 3, x_3 = 4, x_4 = 5, x_5 = 6.$$

- Calculer

$$\sup_{x \in \mathbb{R}} \left| F_0(x) - \widehat{F}(x) \right|$$

On pourra recopier et compléter le tableau suivant

x_i	2.00	3.00	4.00	5.00	6.00
$F(x_i)$					
E_i^+					
E_i^-					
$\max(E_i^+, E_i^-)$					

- Expliquer comment on peut conclure.

Correction Examen Statistique
Mercredi 16 Janvier 2008

Exercice 1 : Estimation

① La vraisemblance des observations s'écrit

$$L(x_1,\dots,x_n;a) = \prod_{i=1}^{n} \frac{a}{x_i^{a+1}} = \frac{a^n}{(\prod x_i)^{a+1}}$$

d'où

$$\ln L(x;a) = n\ln a - (a+1)\sum_{i=1}^{n} \ln x_i \quad \text{avec} \quad x=(x_1,\dots,x_n)$$

$$\frac{\partial \ln L}{\partial a} = \frac{n}{a} - \sum_{i=1}^{n} \ln x_i \geq 0 \iff n \geq a\sum_{i=1}^{n} \ln x_i \iff \boxed{a \leq \frac{n}{\sum_{i=1}^{n} \ln x_i}}$$

On a donc le tableau de variation suivant

a	0		$\frac{n}{\sum \ln x_i}$		$+\infty$
$\frac{\partial \ln L}{\partial a}$		$+$	0	$-$	
$\ln L$		↗		↘	

$\ln L(x;a)$ possède donc un maximum global unique obtenu pour $a = \frac{n}{\sum_{i=1}^{n} \ln x_i}$

d'où

$$\boxed{\hat{a}_{MV} = \frac{n}{\sum_{i=1}^{n} \ln x_i}}$$

- Si X suit une loi de Pareto de paramètre a, alors la densité de $Z = \ln X$ s'écrit

$$g(z) = \frac{a\left|\frac{dx}{dz}\right|}{(e^z)^{a+1}} \qquad z \geq 0$$

car le changement de variables $Z = \ln X$ ($\Leftrightarrow X = e^Z$) est bijectif de $[1,+\infty[$ dans $[0,+\infty[$. Donc

$$g(z) = a\,e^{-(a+1)z}\,e^{z} = a\,e^{-az}, \quad z \geq 0$$

c'est-à-dire $\boxed{g(z) = a\,e^{-az}\,\mathbb{1}_{\mathbb{R}^+}(z)}$

Z suit donc une loi Gamma de paramètres $\theta = a$ et $\nu = 1$, soit $\boxed{Z \sim \Gamma(a,1)}$

- La fonction caractéristique de $U = \sum_{i=1}^{n} Z_i$ s'écrit

$$\Phi_U(t) = E[e^{itU}] = E\left[e^{it\sum_{k=1}^{n} Z_k}\right] = E\left[\prod_{k=1}^{n} e^{itZ_k}\right] \underset{Z_1,\dots,Z_n \text{ ind}}{=} \prod_{k=1}^{n} E[e^{itZ_k}]$$

$$\Phi_U(t) = \prod_{k=1}^{n} \frac{1}{1-i\frac{t}{a}} = \boxed{\frac{1}{\left(1-i\frac{t}{a}\right)^n}}$$

↑ Tables

C'est la fonction caractéristique d'une loi Gamma $\Gamma(a,n)$ donc

$$\boxed{U \sim \Gamma(a,n)}$$

La moyenne et la variance de $\frac{1}{U}$ s'écrivent

$$E\left[\frac{1}{U}\right] = \int_0^{+\infty} \frac{1}{u} \frac{a^n}{\Gamma(n)} e^{-au} u^{n-1} du \overset{v=au}{=} \int_0^{+\infty} \frac{1}{\frac{v}{a}} \frac{a^n}{\Gamma(n)} e^{-v} \frac{v^{n-1}}{a^{n-1}} \frac{dv}{a}$$

d'où

$$E\left[\frac{1}{U}\right] = \frac{1}{\Gamma(n)} a \int_0^{+\infty} e^{-v} v^{n-2} dv = \frac{a}{\Gamma(n)} \Gamma(n-1) = \boxed{\frac{a}{n-1}}$$

$$\mathrm{Var}\left[\frac{1}{U}\right] = V\left[\frac{1}{U}\right] = E\left[\frac{1}{U^2}\right] - E\left[\frac{1}{U}\right]^2$$

$$E\left[\frac{1}{U^2}\right] = \int_0^{+\infty} \frac{1}{u^2} \frac{a^n}{\Gamma(n)} e^{-au} u^{n-1} du \overset{v=au}{=} \int_0^{+\infty} \frac{a^n}{\Gamma(n)} e^{-v} \frac{v^{n-3}}{a^{n-3}} \frac{dv}{a}$$

$$= \frac{a^2}{\Gamma(n)} \int_0^{+\infty} e^{-v} v^{n-3} dv = a^2 \frac{\Gamma(n-2)}{\Gamma(n)} = \boxed{\frac{a^2}{(n-1)(n-2)}}$$

$$\mathrm{Var}\left[\frac{1}{U}\right] = \frac{a^2}{(n-1)(n-2)} - \frac{a^2}{(n-1)^2} = \frac{a^2}{(n-1)} \frac{n-1-(n-2)}{(n-1)(n-2)} = \boxed{\frac{a^2}{(n-1)^2(n-2)}}$$

L'estimateur du maximum de vraisemblance s'écrit

$$\hat{a}_{MV} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{n}{U}$$

donc $E[\hat{a}_{MV}] = n\,E\left[\frac{1}{U}\right] = \boxed{\frac{n}{n-1} a}$ — Un estimateur non biaisé du paramètre a est donc

$$\tilde{a} = \frac{n-1}{n} \hat{a}_{MV} = \boxed{\frac{n-1}{\sum_{i=1}^{n} \ln x_i}}$$

La variance de $\tilde{a}$ s'écrit

$$\mathrm{Var}\,\tilde{a} = \left(\frac{n-1}{n}\right)^2 \mathrm{var}\left(\frac{n}{U}\right) = \left(\frac{n-1}{n}\right)^2 n^2 \frac{a^2}{(n-1)^2(n-2)} = \boxed{\frac{a^2}{n-2}}$$

La borne de Cramer-Rao associée à un estimateur non biaisé de a est

$$BCR(a) = \frac{-1}{E\left[\frac{\partial^2 \ln L(x;a)}{\partial a^2}\right]}$$

mais $\frac{\partial^2 \ln L(x;a)}{\partial a^2} = -\frac{n}{a^2}$ d'où $\boxed{BCR(a) = \frac{a^2}{n}}$

Puisque $\text{var}\,\tilde{a} \neq BCR(a) = \frac{a^2}{n}$, $\tilde{a}$ n'est pas l'estimateur efficace du paramètre a. Par contre, $\hat{a}$ est un estimateur asymptotiquement efficace de a

(2) La densité a posteriori du paramètre a s'écrit

$$f(a|x) = f(a|x_1,\dots,x_n) = \frac{f(x_1,\dots,x_n|a)\, g(a)}{f(x_1,\dots,x_n)}$$

$$\propto f(x_1,\dots,x_n|a)\, g(a)$$

donc $f(a|x) \propto \frac{a^n}{\underbrace{(\prod x_i)^{a+1}}_{e^{(a+1)\sum_{i=1}^n \ln x_i}}} \lambda e^{-\lambda a} = \lambda a^n e^{-a\left[\lambda + \sum_{i=1}^n \ln x_i\right]} e^{-\sum_{i=1}^n \ln x_i}$

$$\propto \boxed{a^n e^{-a\left[\lambda + \sum_{i=1}^n \ln x_i\right]}}$$

soit $\boxed{a|X_1,\dots,X_n \sim \Gamma^2\left(\lambda + \sum_{i=1}^n \ln x_i,\ n+1\right)}$

L'estimateur MAP du paramètre a maximise le logarithme de la densité a Posteriori

$$\ln f(a|x) = n \ln a - a\left[\lambda + \sum_{i=1}^n \ln x_i\right]$$

$$\frac{\partial \ln f(a|x)}{\partial a} \geq 0 \iff \frac{n}{a} \geq \lambda + \sum_{i=1}^n \ln x_i$$

$$\iff a \leq \frac{n}{\lambda + \sum_{i=1}^n \ln x_i}$$

d'où le tableau de variation

a	0		$\frac{n}{\lambda + \sum \ln x_i}$		$+\infty$
$\frac{\partial \ln f(a\|x)}{\partial a}$		$+$	0	$-$	
		$\nearrow$		$\searrow$	

d'où l'estimateur MAP de a:

$$\hat{a}_{MAP} = \frac{n}{\alpha + \sum_{i=1}^{n} \ln x_i} = \frac{1}{\frac{\alpha}{n} + \frac{1}{n}\sum_{i=1}^{n} \ln x_i}$$

quand n est "grand", $\frac{\alpha}{n} \ll \frac{1}{n}\sum_{i=1}^{n} \ln x_i$ donc $\hat{a}_{MAP}$ se comporte comme $\hat{a}_{MV}$

L'estimateur MMSE du paramètre a s'écrit

$$\hat{a}_{MMSE} = \underbrace{E[a \mid x_1, \dots, x_n]}_{\text{c'est la moyenne d'une loi Gamma } (\alpha + \sum_{i=1}^{n} \ln x_i, n+1)} = \frac{n+1}{\alpha + \sum_{i=1}^{n} \ln x_i}$$

Exercice 2 : tests statistiques

1) Le test de Neyman-Pearson pour le problème considéré s'écrit

Reget de H_0 si $\frac{L(x_1, \dots, x_n \mid H_1)}{L(x_1, \dots, x_n \mid H_0)} > k_\alpha$

Soit Reget de H_0 si $\frac{a_1^{n}}{(\prod x_i)^{a_1+1}} \frac{(\prod x_i)^{a_0+1}}{a_0^{n}} > k_\alpha$

si $n \ln\left(\frac{a_1}{a_0}\right) + \underbrace{(a_0+1-a_1-1)}_{a_0 - a_1 < 0} \sum_{i=1}^{n} \ln x_i > \ln k_\alpha$

Soit Reget de H_0 si $\boxed{U(x_1, \dots, x_n) \triangleq U = \sum_{i=1}^{n} \ln x_i < s_\alpha}$

- on a

$$\alpha = P[\text{Reject } H_0 \mid H_0 \text{ vraie}] = P[U < s_\alpha \mid a = a_0]$$

$$\alpha = P[U < s_\alpha \mid U \sim \Gamma(a_0, n)]$$

Soit $\boxed{\alpha = \int_0^{s_\alpha} \frac{a_0^{n}}{\Gamma(n)} e^{-a_0 u} u^{n-1} du} \quad (*)$

En faisant le changement de variables $v = a_0 u$, on obtient

$$\alpha = \int_0^{a_0 s_\alpha} \frac{a_0^{n}}{\Gamma(n)} e^{-v} \frac{v^{n-1}}{a_0^{n-1}} \frac{dv}{a_0} = \frac{1}{\Gamma(n)} \int_0^{a_0 s_\alpha} e^{-v} v^{n-1} dv$$

On a donc $\alpha = I_n(a_0 S_\alpha)$

d'où $$\boxed{S_\alpha = \frac{1}{a_0} I_n^{-1}(\alpha)}$$

- De même la puissance du test s'écrit

$$\Pi = P[\text{Rejeter } H_0 \mid H_1 \text{ vraie}] = P[U < S_\alpha \mid U \sim \Gamma(a_1, n)]$$

Soit

$$\Pi = \int_0^{S_\alpha} \frac{a_1^n}{\Gamma(n)} e^{-a_1 u} u^{n-1} du$$

En faisant le changement de variables $v = a_1 u$, on obtient

$$\Pi = \int_0^{a_1 S_\alpha} \frac{a_1^n}{\Gamma(n)} e^{-v} \frac{v^{n-1}}{a_1^{n-1}} \frac{dv}{a_1} = \int_0^{a_1 S_\alpha} \frac{e^{-v} v^{n-1}}{\Gamma(n)} dv$$

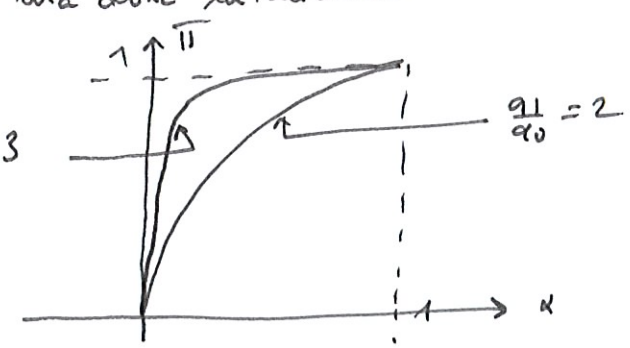
c'est-à-dire $$\boxed{\Pi = \alpha I_n(a_1 S_\alpha)}$$

- Les courbes COR du test sont définies par les relations entre Π et α, c'est-à-dire

$$\Pi = I_n\left(a_1 \frac{1}{a_0} I_n^{-1}(\alpha)\right)$$

d'où $$\boxed{\Pi = I_n\left(\frac{a_1}{a_0} I_n^{-1}(\alpha)\right)}$$

On voit donc que la performance du test ne dépend que du rapport $\frac{a_1}{a_0}$. Le test est d'autant meilleur que $\frac{a_1}{a_0}$ est grand. Donc le deuxièmes cas $(a_0, a_1) = (1,3)$ donc $\frac{a_1}{a_0} = 3$ tandis que dans le premier cas $(a_0 a_1) = (1,2)$, on a $\frac{a_1}{a_0} = 2$. Le deuxième cas donnera donc de meilleures performances. L'allure des courbes COR sera donc la suivante :



2) La fonction de répartition d'une loi de Pareto de paramètre a s'écrit

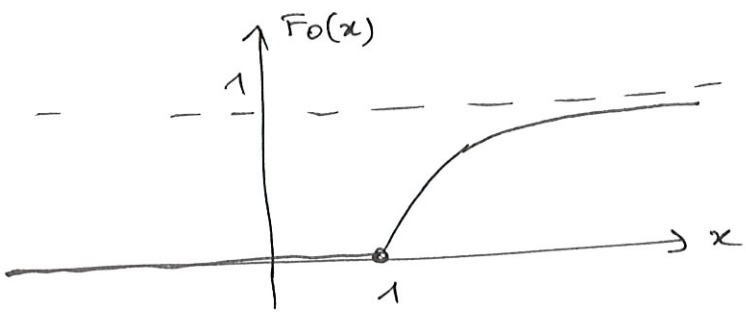
$$F_0(x) = P[X < x] = \begin{cases} 0 & \text{si } x < 1 \\ \int_1^x \frac{a}{u^{a+1}} du & \text{si } x \geq 1 \end{cases}$$

Pour $x \geq 1$, on obtient donc $F_0(x) = \int_1^x a u^{-a-1} du$

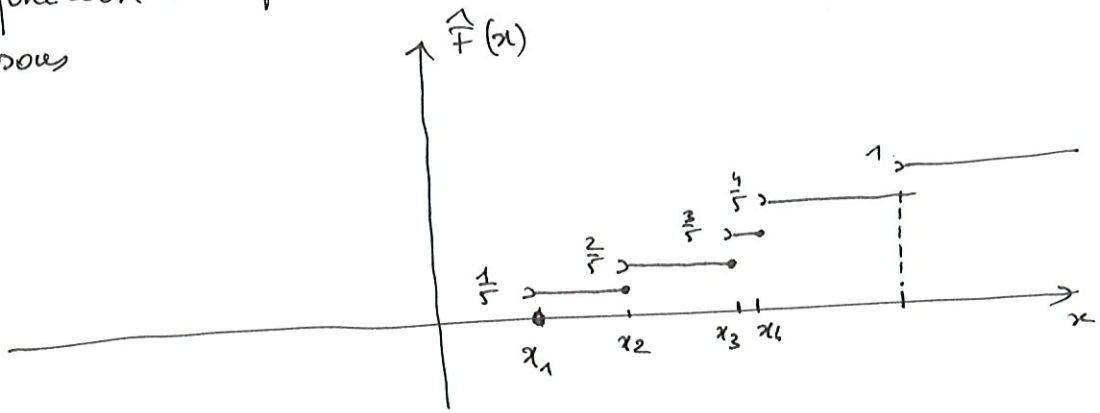
$$= \left[-u^{-a}\right]_1^x = 1 - x^{-a}$$

donc $$\boxed{F_0(x) = \left[1 - \frac{1}{x^a}\right] \mathbb{1}_{[1,+\infty[}(x)}$$

Pour $a=1$, on a $F_0(x) = \left[1 - \frac{1}{x}\right] \mathbb{1}_{[1,+\infty[}(x)$ qui est représentée ci-dessous



- La fonction de répartition des données ~~d'ordre~~ est représentée ci dessous



- on sait que $\sup_{x \in \mathbb{R}} |F_0(x) - \hat{F}(x)| = \sup_{i \in \{1,\dots,5\}} \text{Max}\left(E_i^+, E_i^-\right)$

avec $E_i^+ = \left|F_0(x_i) - \frac{i}{n}\right|$ $E_i^- = \left|F_0(x_i) - \frac{i-1}{n}\right|$ avec $n=5$

c'est-à-dire, dans le cas présent

$F_0(x_1) = \frac{1}{2}$ $F_0(x_2) = 1 - \frac{1}{3} = \frac{2}{3}$ $F_0(x_3) = 1 - \frac{1}{4} = \frac{3}{4}$ $F_0(x_4) = 1 - \frac{1}{5} = \frac{4}{5}$

$F_0(x_5) = 1 - \frac{1}{6} = \frac{5}{6}$

donc

$$\sup_{x\in\mathbb{R}} |F_0(x)-\hat{F}(x)| = \max\left\{E_1^-, E_1^+, E_2^-, E_2^+, \dots, E_5^-, E_5^+\right\}$$

$x_1 = 2 \Rightarrow$ $E_1^+ = \left|\frac{1}{2} - \frac{1}{5}\right| = \frac{3}{10} = 0.3$

$E_1^- = \left|\frac{1}{2} - 0\right| = \frac{1}{2} = 0.5$

$x_2 = 3 \Rightarrow$ $E_2^+ = \left|\frac{2}{3} - \frac{2}{5}\right| = \frac{4}{15} = 0.267$

$E_2^- = \left|\frac{2}{3} - \frac{1}{5}\right| = \frac{7}{15} \# 0.467$

$x_3 = 4 \Rightarrow$ $E_3^+ = \left|\frac{3}{4} - \frac{3}{5}\right| = \frac{3}{20} = 0.150$

$E_3^- = \left|\frac{3}{4} - \frac{2}{5}\right| = \frac{7}{20} = 0.350$

$x_4 = 5 \Rightarrow$ $E_4^+ = \left|\frac{4}{5} - \frac{4}{5}\right| = 0$

$E_4^- = \left|\frac{4}{5} - \frac{3}{5}\right| = \frac{1}{5} = 0.2$

$x_5 = 6 \Rightarrow$ $E_5^+ = \left|\frac{5}{6} - 1\right| = \frac{1}{6} = 0.167$

$E_5^- = \left|\frac{5}{6} - \frac{4}{5}\right| = \frac{1}{30} = 0.033$

d'où

$$\boxed{\sup_{x\in\mathbb{R}} |F_0(x) - \hat{F}(x)| = \frac{1}{2} = 0.5}$$

- Pour conclure, on se fixe un risque α, par exemple $\alpha = 0.01$ ou $\alpha = 0.05$ on regarde l'erreur admissible donnée par Kolmogorov et on

accepte H_0 si $\sup_{x\in\mathbb{R}} |F_0(x) - \hat{F}(x)|$ est inférieur à cette erreur

On n'oubliera pas de donner une réponse de la forme " on rejette H_0 avec le risque $\alpha = 0.01$ "