

# Projet STA 201

Travail élaboré par :

Khalil HARRABI

Maha OUALI

## Introduction

La loi de Pareto tire son nom de son inventeur l'économiste et sociologue italien Vilfredo Pareto.

Cette loi possède des domaines d'applications multiples à savoir l'économie, stratégie commerciale, gestion des ressources humaines... .

Dans ce projet, on vise à faire une étude théorique aussi qu'une simulation numérique de cette loi.

On essaie ensuite de tester si on peut appliquer cette loi pour estimer les taux de naissances dans les grandes villes d'un pays.

## Partie I:

1)

Pour que l'espérance soit finie il faut que  $x f_{\beta}^c(x)$  soit intégrable sur  $[\beta, +\infty]$   
donc il faut que  $\frac{c}{x^2}$  soit intégrable sur  $[\beta, +\infty]$   
Alors c'est  $c > 1$

On effectuant le même raisonnement il faut que  
 $\frac{c}{x^2}$  soit intégrable sur  $[\beta, +\infty]$   
c'est  $c > 2$

2) On commence par calculer

$$E(X_1) = \int_{\beta}^{+\infty} x f_{\beta}^c(x) dx$$

$$= \int_{\beta}^{+\infty} \frac{c}{x^2} dx = \left[ -\frac{c}{x} \right]_{\beta}^{+\infty}$$

$$\underline{E(X_1) = \frac{c}{c-1}}$$

$$\Leftrightarrow c = \frac{E(X)}{E(X)-1}$$

On pose  $\hat{m}_1$  l'estimateur de moment d'ordre 1

$$\text{donc } \hat{c} = \frac{\hat{m}_1}{\hat{m}_1 - 1}$$

est l'estimateur par la méthode des moments de  $c$

(1)

consistance pour  $z > 1$

Par passage à la ~~mo~~ limite

$$\hat{z} \xrightarrow{\text{ }} \frac{E(x_n)}{E(x_n) - 1} = \frac{\frac{z}{z-1}}{\frac{z}{z-1} - 1} = z$$

donc l'estimateur est consistant.

3)

$$f_z(x) = \frac{z}{x^{z+1}} = \exp(\ln(z) - (z+1)\ln(x))$$

On peut l'écrire sous la forme

$$\exp(a(x)\alpha(z) + \beta(z) + c(x))$$

avec par identification

$$a(x) = \ln(x) \quad \alpha(z) = -(z+1)$$

$$c(x) = 0 \quad \beta(z) = \ln(z)$$

donc la loi  $P_a(z)$  appartient à une famille exponentielle.

La statistique exhaustive est

$$T_m = \sum_{i=1}^m \ln(x_i)$$

\* Montrons que  $\tau \sum \log(x_i)$  suit une loi gamma  $G_\alpha(n, 1)$

$$P(\log(x_i) \leq t) = P(x_i \leq \exp(t)), \quad \exp(t) > 1$$

$$= \int_0^{\exp(t)} \frac{z}{x^{z+1}} dx = z \left[ \frac{-1}{2} \frac{1}{x^2} \right]_0^{\exp(t)}$$

$$= 1 - \frac{1}{\exp(zt)} = 1 - \exp(-zt)$$

donc  $\log(x_i) \sim \exp(\frac{1}{z})$   
alors

$$z \sum \log(x_i) \sim \text{Ga}(m, 1)$$

Les variables  $(x_i)_{i \in \mathbb{N}}$  sont iid d'espérance finie et de variance finie donc on peut appliquer la LGN + TCL.

4)

$$\begin{aligned} L(z, x_1, \dots, x_m) &= \prod_{i=1}^m f_z(x_i) = \prod_{i=1}^m \frac{z}{(x_i)^{z+1}} \\ &= \frac{z^m}{\left(\prod_{i=1}^m x_i\right)^{z+1}} \end{aligned}$$

$$\log(L(z, x_1, \dots, x_m)) = m \log z - (z+1) \sum_{i=1}^m \log(x_i)$$

$$\frac{\partial \log L}{\partial z} = \frac{m}{z} - \sum_{i=1}^m \log(x_i) = 0$$

$$z = \frac{m}{\sum_{i=1}^m \log(x_i)}$$

$$\text{De plus, } \frac{\partial^2 \log L}{\partial z^2} = -\frac{m}{z^2} < 0$$

$$\text{Ainsi } \hat{z} = \frac{m}{\sum_{i=1}^m \log(x_i)} \text{ est l'EIV}$$

Cherchons l'information de Fisher :

$$I_m(z) = -E\left(\frac{\partial^2 \log L(z, x_1, \dots, x_m)}{\partial z^2}\right) = \frac{m}{z^2}$$

$$\text{Var}(\hat{\tau}) = \text{Var}\left(\frac{m\bar{\tau}}{\bar{\tau} \sum_{i=1}^n \log(x_i)}\right) = (m\bar{\tau})^2 \text{Var}\left(\frac{1}{u}\right)$$

$$\text{avec } u = \bar{\tau} \sum_i \log(x_i)$$

On pose

$$f: u \mapsto \frac{1}{u}$$

$$\text{Var}(\hat{\tau}) = (m\bar{\tau})^2 \text{Var}(f(u))$$

$$= (m\bar{\tau})^2 \left[ E(f(u)^2) - [E(f(u))]^2 \right]$$

$$E(f(u)^2) = \int_0^\infty \frac{1}{x^2} \frac{x^{m-1} \exp(-x)}{\Gamma(m)} dx$$

$$\stackrel{\text{IPP}}{=} \frac{1}{m-2} + \frac{1}{m-1}$$

$$E(f(u))^2 = \frac{1}{(m-1)^2}$$

$$\text{Var}(\hat{\tau}) = (m\bar{\tau})^2 \left( \frac{1}{(m-1)(m-2)} - \frac{1}{(m-1)^2} \right) \neq I_m(\tau)^{-1}$$

donc  $\hat{\tau}$  n'est pas efficace.

(On peut conclure directement la non efficacité de l'estimation par le fait qu'il n'est pas consistant)

$$\star \quad \hat{\tau} = \frac{m}{\sum_{i=1}^m \log(x_i)} = \frac{m}{\bar{\tau}}$$

comme les  $(x_i)_{i \in \mathbb{N}}$  sont iid alors  $(\log(x_i))_{i \in \mathbb{N}}$  sont iid

④

On applique le TCI

$$\sqrt{m} \frac{\sum_{i=1}^m \log(x_i) - \frac{1}{\tau}}{\frac{1}{\tau}} \rightsquigarrow N(0, 1)$$

On applique la delta-méthode

$$\text{avec } h: x \mapsto \frac{1}{x}$$

$$\hat{\tau} = h\left(\frac{1}{m} \sum_{i=1}^m \log(x_i)\right)$$

On obtient:

$$\sqrt{m} (\hat{\tau} - \tau) \rightsquigarrow N(0, \tau^2)$$

On retrouve donc la normalité asymptotique de l'EMV.

5)

$$H_0: \tau \leq \tau_0 = 3$$

$$H_1: \tau > \tau_0$$

Test de Neymann-Pearson :

$$RV = \frac{L(\tau_0, x)}{L(\tau_1, x)} = \left(\frac{\tau_0}{\tau_1}\right)^m \left(\prod_{i=1}^m X_i\right)^{\tau_0 - \tau_1}$$

$$= \left(\frac{\tau_0}{\tau_1}\right)^m \exp((\tau_0 - \tau_1) \sum \log X_i)$$

Ainsi RV est une fonction décroissante de la statistique exhaustive  $T = \sum_i \log X_i$  pour  $\tau_1 > \tau_0$ .

i) On a de plus sous  $H_0$   $\tau \leq \sum_{i=1}^m \log(X_i) \sim G_{\alpha}(m, 1)$

En appliquant Lehman,

on obtient la zone de Rejet

$$R = \left\{ \tau_0 \sum_{i=1}^m \log(X_i) < q_{\alpha}^{G(m, 1)} \right\}$$

⑤

b) On réécrit le test sous la forme

$$H_0 : A\theta = A\theta_0$$

$$H_1 : A\theta > A\theta_0$$

$$\text{avec } A = [\mathbb{I}] \quad , \quad \theta = [\bar{z}] \text{ et } \theta_0 = [\bar{z}_0]$$

et puisque

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\text{d.f.}} \mathcal{N}(0, 1)$$

car EMV est régulier

$$\text{Donc } W = \sqrt{n}(\bar{z} - \bar{z}_0) \xrightarrow{\text{d.f.}} \mathcal{N}(0, 1)$$

Donc la zone de rejet est

$R_x = \{x, W > q_{1-\alpha}^*\}$  de niveau asymptotique  $\alpha$   
avec  $q_{1-\alpha}^*$  quantile d'ordre  $1-\alpha$  de la loi  $\mathcal{N}(0, 1)$

6) La fonction puissance du test Neyman - Pearson

$$\Pi(x) = P_{\bar{z} > \bar{z}_0} (\bar{z} < T(x) < q_{\alpha}^{Gal(n, 1)})$$

$$= P_{\bar{z} > \bar{z}_0} \left( \bar{z} \sum_{i=1}^n \log(x_i) < \frac{1}{\bar{z}_0} q_{\alpha}^{Gal(n, 1)} \right)$$

$$= F_{Gal(n, 1)} \left( \frac{1}{\bar{z}_0} q_{\alpha}^{Gal(n, 1)} \right)$$

7) Test bilatéral  $H_0 : \bar{z} = \bar{z}_0 = 3$  vs  $H_1 : \bar{z} \neq \bar{z}_0$

$$\text{Cas 1)} \quad R_{\text{sym}}^{\text{gal}} = \left\{ \bar{z}_0 \sum_{i=1}^n \log(x_i) < q_{\frac{\alpha}{2}}^{Gal(n, 1)} \right\} \cup \left\{ \bar{z}_0 \sum_{i=1}^n \log(x_i) > q_{1-\frac{\alpha}{2}}^{Gal(n, 1)} \right\}$$

Cas 2) région de rejet du test biaisé

Zone de rejet

$$R = \left\{ \bar{z}_0 \sum_{i=1}^n \log(x_i) < q_1 \right\} \cup \left\{ \bar{z}_0 \sum_{i=1}^n \log(x_i) > q_2 \right\}$$

avec  $q_1$  et  $q_2$  vérifiant les équations suivantes :

$$\star P_{\tau_0} (q_1 \leq \tau_0 \sum_{i=1}^n \log(x_i) \leq q_2) = 1 - \alpha$$

$$\Leftrightarrow F(q_2) - F(q_1) = 1 - \alpha$$

\* Puisque le test est sans biais, alors sa puissance doit être minimale en  $\tau_0$

$$\Pi(\tau) = 1 - P_{\tau} (q_1 \leq \tau \sum_{i=1}^n \log(x_i) \leq q_2)$$

$$= 1 - P_{\tau} \left( \frac{\tau}{\tau_0} q_1 \leq \tau \sum_{i=1}^n \log(x_i) \leq \frac{\tau}{\tau_0} q_2 \right)$$

$$= 1 - F\left(\frac{\tau}{\tau_0} q_2\right) + F\left(\frac{\tau}{\tau_0} q_1\right)$$

On dérive on obtient :

$$\Pi'(\tau) = -\frac{q_2}{\tau_0} F'\left(\frac{\tau}{\tau_0} q_2\right) + \frac{q_1}{\tau_0} F'\left(\frac{\tau}{\tau_0} q_1\right)$$

$$\Pi'(\tau_0) = 0 \Rightarrow$$

alors

$$q_1 F'(q_1) = q_2 F'(q_2) \Leftrightarrow q_1^n e^{-q_1} = q_2^n e^{-q_2}$$

## Partie 2 :

1) Soit  $U$  une loi uniforme entre  $0$  et  $1$ .

Soit  $g$  une fonction bornée mesurable à valeurs réelles

$$\mathbb{E}(g(u^{-\frac{1}{\alpha}})) = \int_0^1 g(u^{-\frac{1}{\alpha}})$$

$$= \int_{-\infty}^1 g(v) \cdot v^{-\frac{1}{\alpha}-1} dv = \int_1^{+\infty} g(v) \cdot v \cdot \frac{1}{v^{\frac{1}{\alpha}+1}} dv$$

$$= \int_{\mathbb{R}} g(v) \cdot \frac{1}{v^{\frac{1}{\alpha}+1}} \mathbb{1}_{v \geq 1} dv$$

(On a effectué le changement de variable  $v = u^{-\frac{1}{\alpha}}$

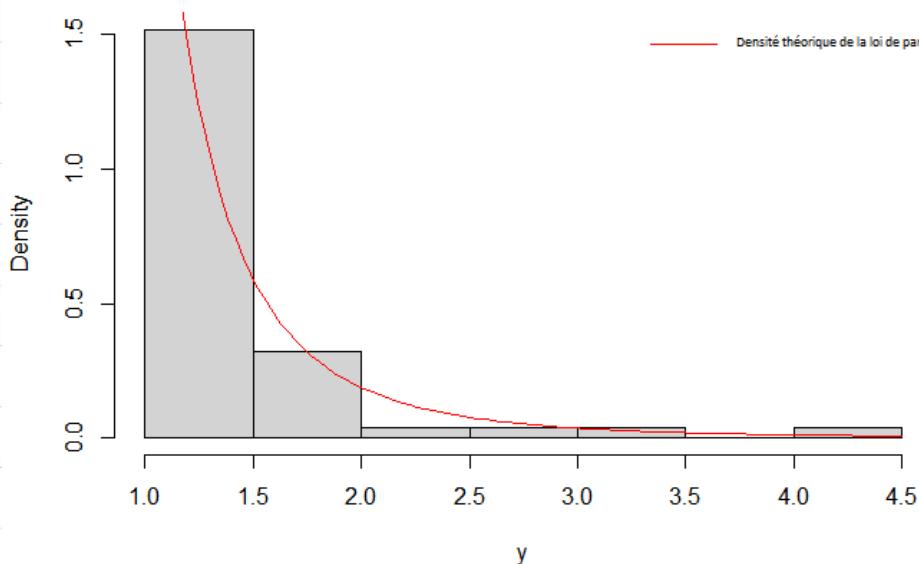
$$\Leftrightarrow u = v^{-\alpha} \text{ et } du = -\alpha v^{-\alpha-1} dv$$

alors

REVV

On génère un échantillon de loi  $\text{Pareto}(c=3)$  de taille  $n=50$  et on lui superpose la courbe de densité on trouve la figure suivante :

Histogramme d'un échantillon  $n=50$  de variables suivant la loi de pareto



La figure obtenue nous permet de déduire que l'échantillon généré suit bien la loi de Pareto de paramètre  $\tau = 3$

- En guise de prendre des décisions pour chacun des tests pour  $\alpha = 0,05$

On calcule les quantiles on obtient que

$$\left( \tau_0 \sum_{i=1}^m \log(x_i) \simeq 48,1 \right) > \left( q_{\alpha}^{Ga(m,1)} \simeq 38,9 \right)$$

Ainsi on conserve  $H_0$  pour la loi exacte du (5.a)) avec une erreur de second ordre espèce inconnue.

On fait de même pour la loi approchée on obtient :

$$\left( \frac{\sqrt{m}}{\tau_0} (\hat{\tau} - \tau_0) = 0,96 \right) < \left( q_{1-\alpha}^* = 1,64 \right)$$

Donc on conserve  $H_0$  avec une erreur de second espèce inconnue

2)

Dans le but d'estimer  $\hat{\tau}$ , on calcule  $B$  fois l'estimation avec la commande replicate (on constate que la formule à l'intérieur de la fonction est donnée par écriture de  $\hat{\tau}$  déterminée précédemment)

On obtient donc un vecteur des valeurs de  $\hat{\tau}$  correspondant à chaque échantillon généré.

On représente l'histogramme demandé et on lui superpose la fonction de densité de la loi Gamma-inverse de paramètres  $m, m\tau$

( ce choix de densité est justifié par le fait que

$$\tau \sum_{i=1}^m \log x_i \sim Ga(m, 1) \text{ ainsi } \frac{\sum_{i=1}^m \log x_i}{m} \sim Ga\left(m, \frac{1}{m\tau}\right)$$

$$\text{d'où } \hat{\tau} = \frac{m}{\sum_{i=1}^m \log(x_i)} \sim Ga\left(m, m\tau\right)$$

Pour déterminer le nombre moyen de rej. de l'hypothèse nulle on compte le nbre moyen de rejcts du test sur  $B$  simulations indépendantes du test.

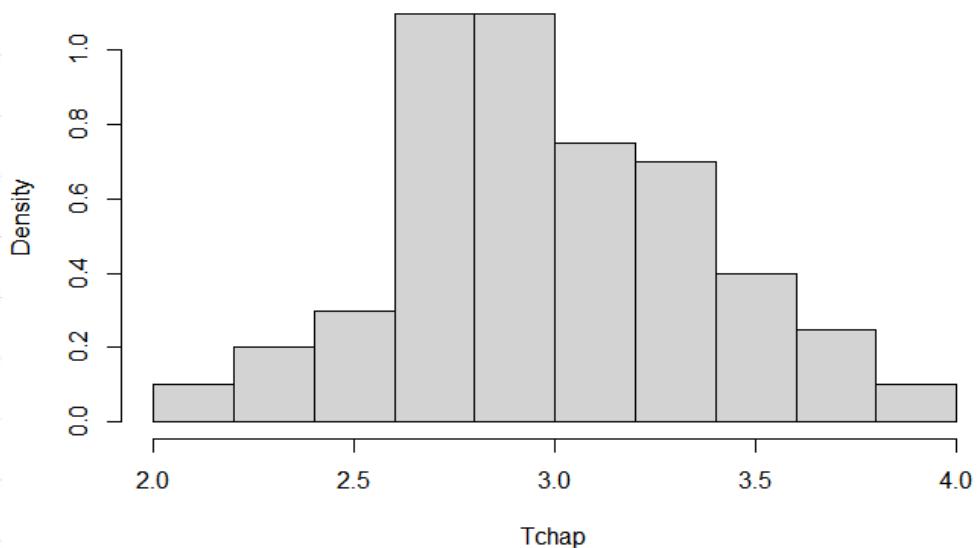
Un calcul numériqu donne une valeur moyenne de nombre de rejcts égale à 0,06 pour le test de Neyman-Pearson et 0,03 pour le test de Wald.

On remarque que la 1<sup>re</sup> valeur est plus proche de  $\alpha$ . Ces écarts sont des conséquences de la génération aléatoire des échantillons.

La 2<sup>ee</sup> valeur est plus éloignée de  $\alpha$  car on a utilisé une loi approchée pour la simulation.

En prenant  $m$  et  $B$  plus grands, on obtiendrait des valeurs de plus en plus proches de  $\alpha$ .

histogramme des simulations des variables suivant la loi inverse-gamma



3)

On trace la courbe de puissance du test unilatéral de la loi exacte  $\tau \rightarrow \Pi(\tau) = F\left(\frac{\tau}{2} q_{\alpha/2}^{G(n,1)}\right)$  quand  $\tau$  varie entre 2 et 5 et on la superpose la droite horizontale horizontale d'ordonnée 0,05.

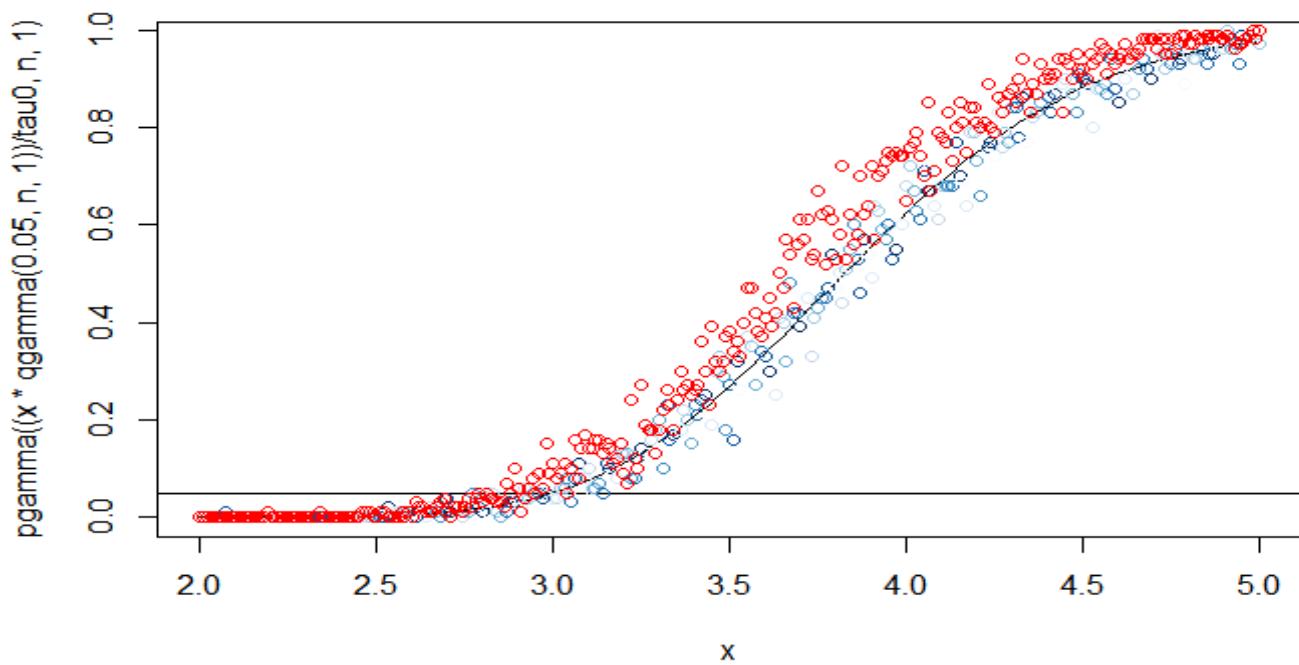
On obtient la figure ci-dessous.

On remarque

On constate d'après cette courbe que : le test est sans biais pour  $\tau > 3$  vu que  $\Pi(\tau) > \alpha$ .

- \* Le test est convergent puisque sa puissance tend vers 1.

courbe de puissance



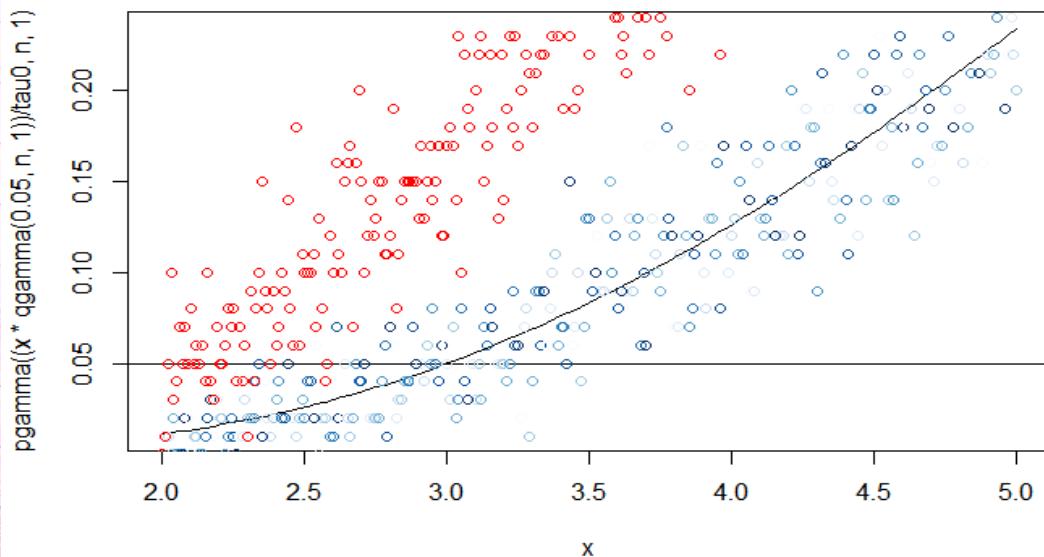
On superpose sur un même graphe les courbes de puissances des tests bilatériques symétriques.

$$\Pi_1(\tau) = 1 - F\left(\frac{\tau}{2} q_{1-\frac{\alpha}{2}}^{G(n,1)}\right) + F\left(\frac{\tau}{2} q_{\frac{\alpha}{2}}^{G(n,1)}\right)$$

est non biaisé.

$\Pi_2(z) = 1 - F\left(\frac{z}{Z_0} q_2\right) + F\left(\frac{z}{Z_0} q_1\right)$  et la droite horizontale d'ordonnée 0,05

courbe de puissance



On remarque que pour  $n=50$  le nuage de points de l'estimation de la puissance entoure la courbe théorique, les estimations et la valeur théorique sont très proches.

Pour  $n=5$ , le nuage des points se éloigne de la courbe théorique.

- Le premier estimateur qui suit la loi gamma reste plus proche précis
- C'est exactement la loi des grands nombres, plus qu'on augmente la taille de l'échantillon plus qu'on s'approche de la valeur réelle.

ii) Dans cette sous-partie, on se propose d'calculer les constantes  $q_1$  et  $q_2$  utilisées pour définir la région de rejet du test bilatéral montrée par

Pour les déterminer on doit résoudre ce système à 2 équations :

$$\begin{cases} F(q_2) - F(q_1) = 1-\alpha \\ q_1^m e^{-q_1} = q_2^m e^{-q_2} \end{cases} \Leftrightarrow \begin{cases} q_2 = F^{-1}(1-\alpha + F(q_1)) \\ q_1^m e^{-q_1} = (F^{-1}(1-\alpha, F(q_1)))^{m-F^{-1}(1-\alpha, F(q_1))} e^{-F^{-1}(1-\alpha, F(q_1))} \end{cases} \quad (1)$$

$$\text{Posons } f : q \mapsto q^m e^{-q} - (F^{-1}(1-\alpha, F(q)))^m e^{-F^{-1}(1-\alpha, F(q))}$$

On trace la courbe de  $f$  pour  $q \in [0, 01, 0, 05]$

et on lui superpose la droite horizontale  $y=0$  (juste pour vérifier)

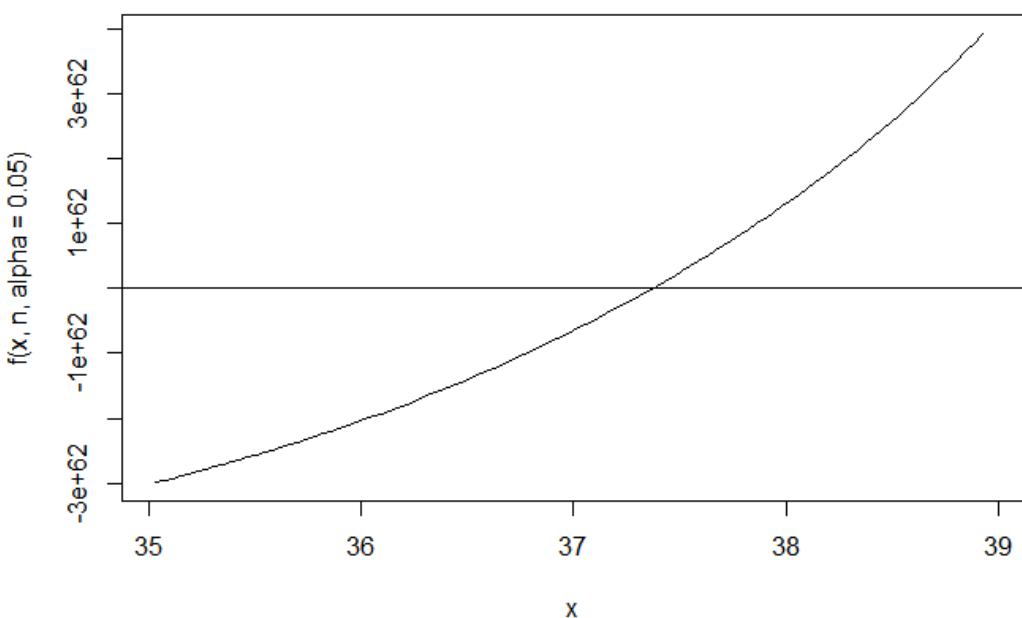
On obtient  $q_1 = 37,3718$

On injecte ce résultat dans (1) on obtient  $q = 65,1955$

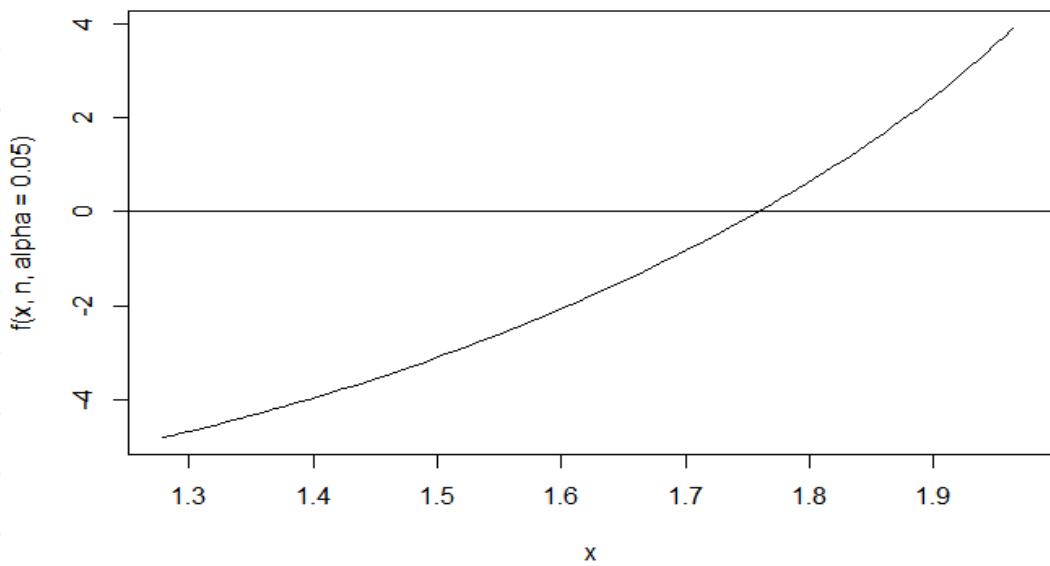
En effectuant la même démarche pour  $m=5$  on trouve

$$q_1 = 1,75808 \text{ et } q_2 = 10,86445$$

### courbe de la fonction $f$ $n=50$



courbe de la fonction  $f$  pour  $n=5$



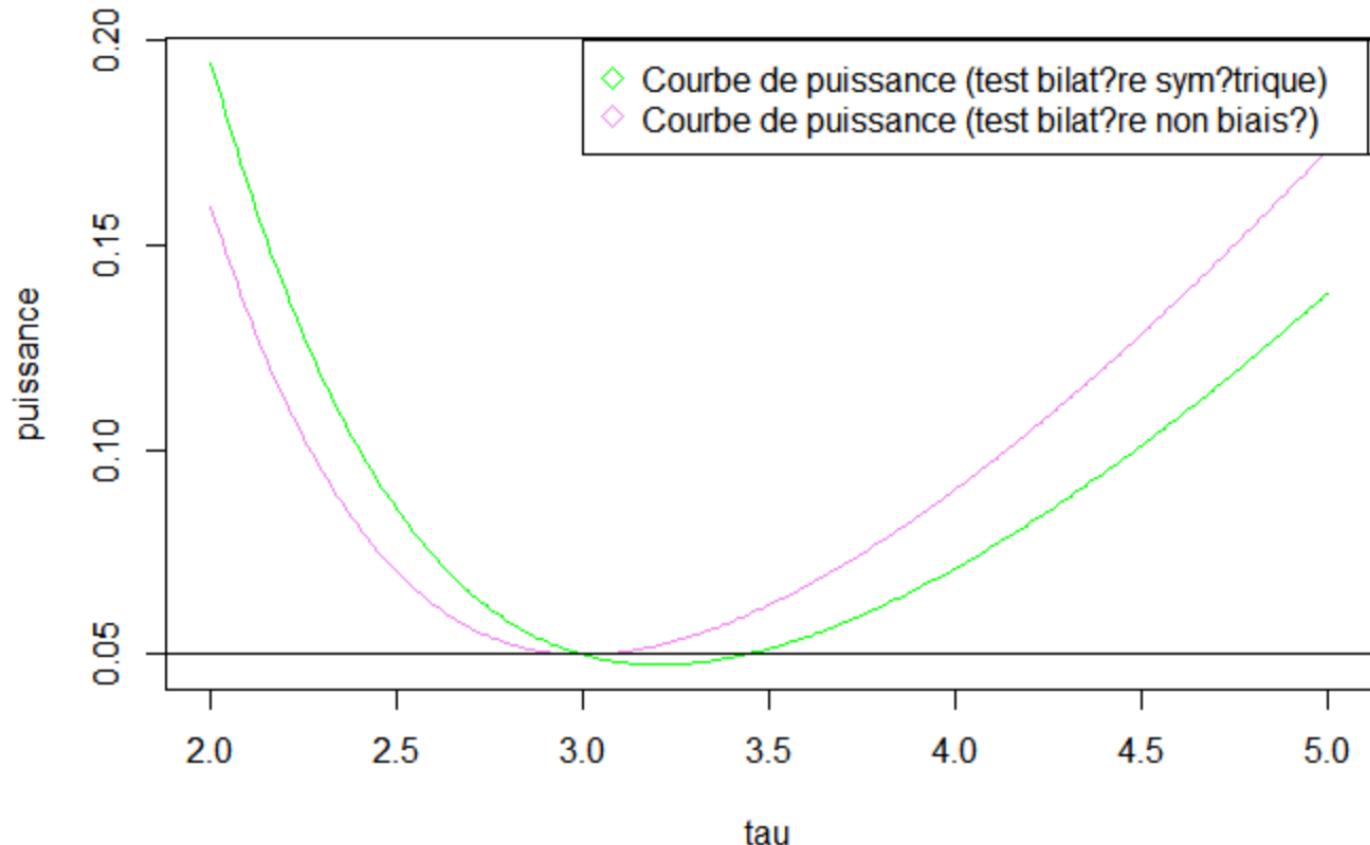
Pour  $n=5$  on va superposer sur un même graph le courbe de puissances des tests bilatères symétriques ( $\alpha_1$ ) et ( $\alpha_2$ ) et son horaire ( $\alpha_2$ ) et on les ajoute la droite horizontale finalisant le niveau.

On vérifie bien d'après la courbe que :

- La courbe de puissance du test non biaisé est : située au dessus de la ligne de niveau  $\alpha = 0,05$  alors que pour  $Z \in [3, 3,5]$  on a  $\Pi(Z) < 0,05$
- pour le test bilatère donc le test est biaisé

On conclut de tout ce qui précède que aucun des deux tests n'est UPP

## courbe de puissance du test bilatéral et du test non biaisé



### Partie 3 :

On considère un échantillon contenant  $m+1$  plus grandes occurrences. Les  $m+1$  statistiques d'ordre

$$X_{[1]} > \dots > X_{[m+1]} \quad (\text{en se rendant de façon décroissante})$$

sont alors observées.

On admet que si  $(x_i)_{i \in [N]}$  est un échantillon de  $N$  variables aléatoires i.i.d. de loi de Pareto  $\text{Pa}(\tau, \beta)$  alors les observations renormalisées  $X_{[1]} / X_{[m+1]} > \dots > \frac{X_{[m]}}{X_{[m+1]}}$  avec  $m \leq N$ ,

correspond au ré-ordonnancement des variables aléatoires de loi de Pareto  $\text{Pa}(\tau)$  de même paramètre  $\tau$  que la loi non renormalisée et ce indépendamment de  $m$  et de  $\beta$ .

#### 1) Les Villes françaises :

En considérant les données de 2018, on normalise les observations en divisant par la valeur minimale et on se débarrasse de cette dernière dans le vecteur obtenu.

On vérifie bien que sa taille est  $m = 39$

On calcule  $\bar{\xi}_F$ :

$$\bar{\xi}_F = 1,818281$$

• Pour pouvoir tracer la courbe des quantiles empiriques en fonction des quantiles d'ordre  $(1:m_F) / (m_F + 1)$

d'une loi  $\text{Pa}(\bar{\xi}_F)$  on doit déterminer la fonction de quantile fonction de répartition de  $\text{Pa}(\tau)$

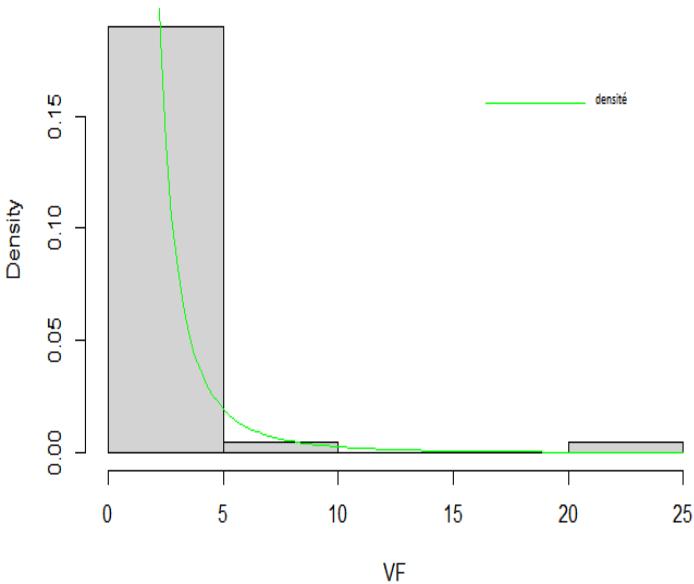
$$F_\tau(x) = \int_a^x \frac{\tau}{t^{\tau+1}} dt = \tau \int_a^x t^{-\tau-1} = 1 - x^{-\tau} \quad \forall x \geq 1$$

$$\text{Soit } q \geq 1 \quad F(q) = y \in [0, 1] \Leftrightarrow 1 - \frac{1}{q^\tau} = y$$

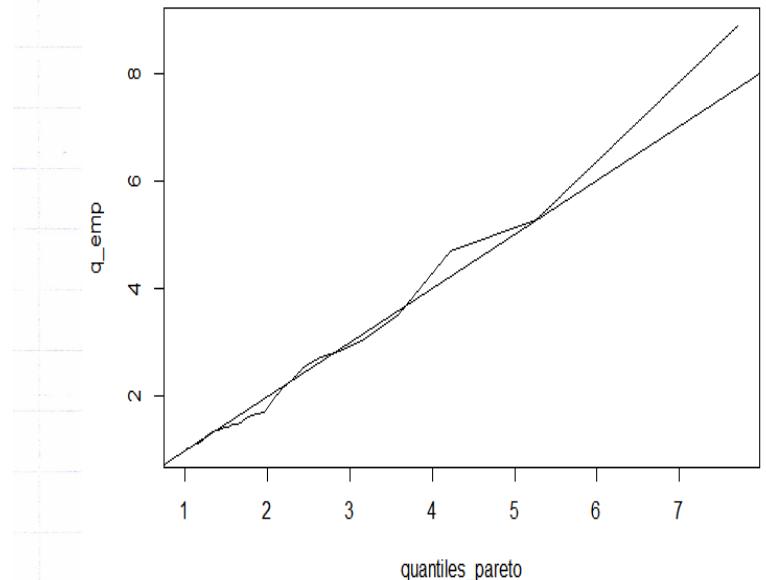
$$\Leftrightarrow q^\tau = \frac{1}{1-y} \Leftrightarrow q = \left( \frac{1}{1-y} \right)^{1/\tau}$$

On superpose le graphique obtenu avec la 1<sup>re</sup> bissectrice  
on obtient la courbe suivante :

histogramme données des villes françaises en 2018



Courbe quantiles empiriques =  $f(\text{quantiles th?oriques de la loi de Pareto})$   
pour les villes en France



Les valeurs obtenues sont proches de la 1<sup>re</sup> bissectrice ce qui permet de justifier qu'on peut approximer la loi des observations par une loi de Pareto.

On applique à cet échantillon le test de Kolmogorov.

Sommaire qui persiste à tester les hypothèses :

H<sub>0</sub>: "Les observations suivent la loi de Pareto de paramètre  $\tau_F$ "

Contre

H<sub>1</sub>: "Les observations ne suivent pas la loi de Pareto"

On cherche - trouve la valeur du p-value = 0,7652 > 0,05

(en utilisant la fonction suggérée)

Ainsi on conserve  $H_0$  avec une erreur de seconde espèce inconnue.

## 2) Villes Allemandes :

On réeffectue la même démarche :

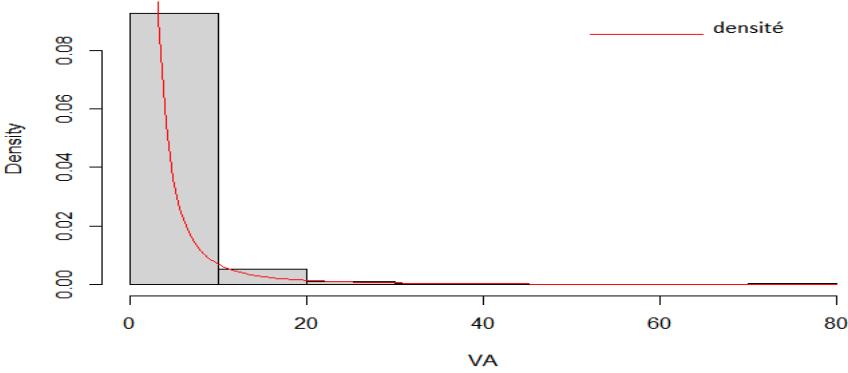
On obtient  $\hat{\zeta}_f = 1,259585$  et la courbe quantile-quantile ci-dessous.

On remarque que l'approximation de la loi de Pareto est vraie pour  $q \in [0, 10]$  avec  $q$  le quantile de loi de Pareto.

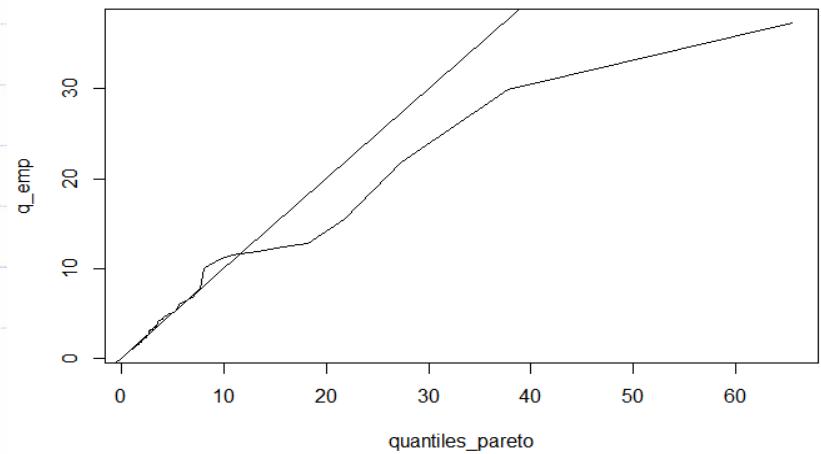
Pour  $q > 10$  la courbe s'éloigne de la 1<sup>re</sup> bissectrice donc le modèle de loi de Pareto est moins adapté.

Le test de Kolmogorov-Smirnov donne  $p\text{-value} = 0,9673 > \alpha$   
⇒ On conserve  $H_0$  avec une erreur de seconde espèce inconnue.

histogramme données des villes allemandes en 2020



Irre quantiles empiriques en fonction des quantiles théoriques de la loi de Pareto pour les villes en Allemagne



### 3) Ville du Royaume Uni

On obtient dans ce cas :  $\hat{T}_F = 1,35875$  et la courbe quantile - quantile ci-dessous :

On remarque que dans ce cas l'approximation est vraie pour  $q \in [0, 5]$

Pour  $q > 5$  la courbe  $q_q$  s'éloigne de la 1<sup>re</sup> bissectrice et on constate qu'elle s'éloigne encore plus que pour les cas précédent

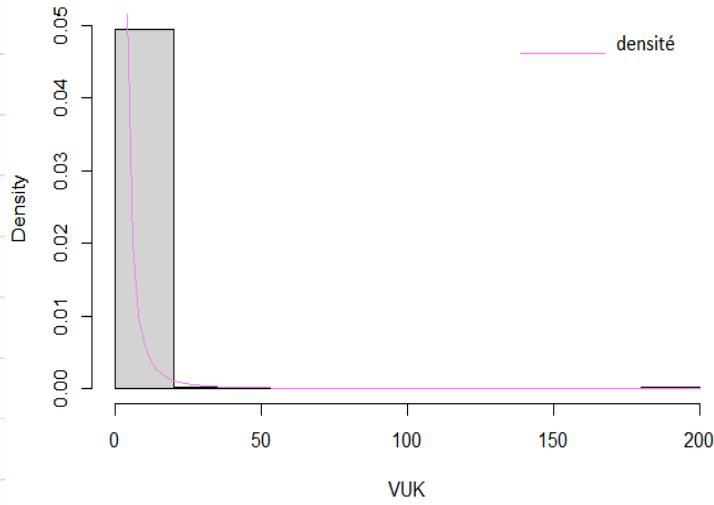
Ainsi l'approximation par la loi de Pareto est

moins adaptée.

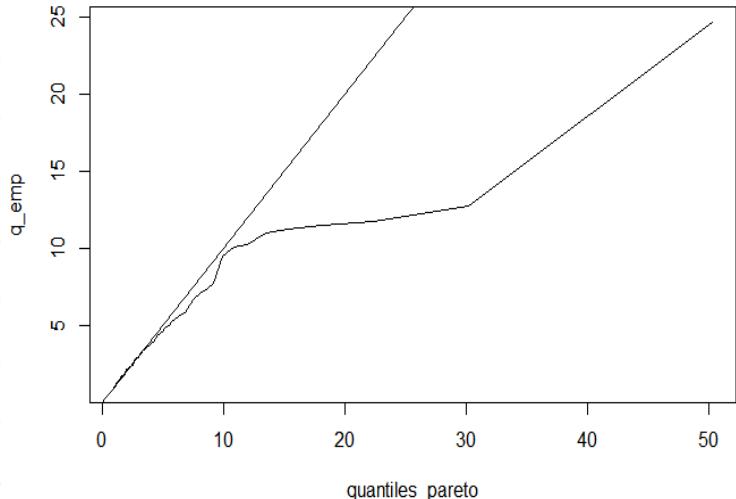
Test de Kolmogorov-Smirnov donne p-value = 0,3211 >  $\alpha$

→ On conserve  $H_0$  avec un seuil de seconde espèce inconnue.

histogramme données des villes UK en 2019



Irre quantiles empiriques en fonction des quantiles théoriques de la loi de Pareto pour les villes de l'UK



## • Conclusions :

Après faire l'étude théorique et la simulation numérique de la loi de Pareto ; on parvient à conclure que le modèle associé à la loi est intéressant pour l'étude et l'estimation de certains phénomènes réels.

Néanmoins, il peut présenter certaines erreurs .  
En effet, pour l'exemple étudié les approximations ne sont vraies que pour certaines valeurs de quantiles et ils sont mal justifiés pour les autres .