TP2

Simon Darnault, Lorenzo Lucas – Roblot

2024-03-07

Contents

Théorème Central Limite et Estimation Monte Carlo

Exercice 7

Soit $\alpha \in \mathbb{R}$. Par définition de la densité de $X \sim \text{Pareto}(a, \alpha)$, notée f, on a :

$$\begin{aligned}
\forall x \in \mathbb{R},\ F_X(x) = \mathbb{P}(X \leqslant x) &= \int_{-\infty}^{x} f(t; a, \alpha)\,\mathrm{d}t && \text{(définition)} \\
&= \int_{-\infty}^{x} \alpha \frac{a^\alpha}{t^{\alpha+1}} \mathbb{1}_{[a,+\infty[}(t)\,\mathrm{d}t && \text{(définition de la densité de la loi de Pareto)} \\
&= \alpha a^\alpha \int_a^x t^{-\alpha-1}\,\mathrm{d}t && \text{(linéarité de l'intégrale et cas x } \geqslant \text{ a)} \\
&= \alpha a^\alpha \left[-\frac{t^{-\alpha}}{\alpha}\right]_{t=a}^{x} \\
&= 1 - \left(\frac{a}{x}\right)^\alpha && \text{(après simplification)}
\end{aligned}$$

On obtient alors la fonction de répartition de la loi de Pareto :

$$\forall x \in \mathbb{R},\ F_X(x) = \left(1 - \left(\frac{a}{x}\right)^\alpha\right) \mathbb{1}_{[a,+\infty[}(x)$$

On suppose ici que $\alpha > 1$. On calcule alors l'espérance théorique d'une loi de Pareto X :

$$\begin{aligned}
\mathbb{E}(X) &= \int_{\mathbb{R}} x f(x; a, x)\,\mathrm{d}x && \text{(définition de l'espérance)} \\
&= \int_{\mathbb{R}} \mathbb{P}(X > x)\,\mathrm{d}x && \text{(propriété de l'espérance pour une variable aléatoire positive)} \\
&= \int_{\mathbb{R}} 1 - \mathbb{P}(X \leqslant x)\,\mathrm{d}x && \text{(événement complémentaire)} \\
&= \int_{\mathbb{R}} 1 - \left(1 - \left(\frac{a}{x}\right)^\alpha\right) \mathbb{1}_{[a,+\infty[}(x)\,\mathrm{d}x && \text{(résultat précédent)} \\
&= \int_0^a \mathrm{d}x + \int_a^{+\infty} \left(\frac{a}{x}\right)^\alpha \mathrm{d}x && \text{(linéarité de l'intégrale)} \\
&= a + a^\alpha \left[\frac{x^{1-\alpha}}{1-\alpha}\right]_{x=a}^{+\infty} \\
&= a + \frac{a}{\alpha - 1} && \text{(calcul de limites)} \\
&= \frac{(\alpha-1)a + a}{\alpha - 1} \\
&= \frac{\alpha a}{\alpha - 1}
\end{aligned}$$

Ainsi :

$$\boxed{\mathbb{E}(X) = \frac{\alpha a}{\alpha - 1}}$$

Nous aurions pu également effectuer le calcul à l'aide de la densité de X :

$$
\begin{aligned}
\mathbb{E}(X) &= \int_{\mathbb{R}} x f(x;a,x)\,\mathrm{d}x && \text{(définition de l'espérance)}\\
&= \int_{\mathbb{R}} \alpha \frac{a^\alpha}{x^\alpha} \mathbb{1}_{[a,+\infty[}\,\mathrm{d}x && \text{(définition de la densité de la loi de Pareto)}\\
&= \alpha a^\alpha \int_a^{+\infty} x^{-\alpha}\,\mathrm{d}x && \text{(linéarité de l'intégrale)}\\
&= \alpha a^\alpha \left[\frac{x^{1-\alpha}}{1-\alpha}\right]_{x=a}^{+\infty}\\
&= \begin{cases} \boxed{+\infty} & \text{si } \alpha \leqslant 1 \\ \frac{\alpha a^\alpha}{\alpha-1} a^{1-\alpha} = \boxed{\frac{\alpha a}{\alpha-1}} & \text{si } \alpha > 1 \end{cases} && \text{(calcul de limites)}
\end{aligned}
$$

Exercice 8

Pour simuler la loi de Pareto, nous avons téléchargé le package `EnvStats`.

```
library(EnvStats)
```

```
##
## Attaching package: 'EnvStats'
```

```
## The following objects are masked from 'package:stats':
##
##     predict, predict.lm
```

Nous noterons $B = 500$ le nombre d'échantillons i.i.d de loi commune Pareto(a, α) de taille n avec $n \in \{20, 100, 200\}$.

```
B <- 500
```

La question précédente assure qu'il est pertinent de prendre $\alpha > 1$. Nous choisirons alors, arbitrairement, $a = 4$ et $\alpha = 5$.

```
a <- 4
alpha <- 5
```

De ce fait, on note $(X_{j,i})_{(j,i)}$ la data frame répertoriant i échantillons de taille n.

```
# Pareto : 500 échantillons de taille n = 20
n <- 20
X <- data.frame(matrix(nrow = n, ncol = B))

# Stockage des simulations
for(i in 1:B) {
  X[,i] <- rpareto(n,a,alpha) # n Pareto(a=4,alpha=5)
}

head(X[1:5,1:8]) # Affiche le début des 8 premiers échantillons
```

```
##         X1       X2       X3       X4       X5       X6       X7       X8
## 1 4.373698 4.353211 4.313648 5.600547 4.562433 4.027642 4.037579 5.058894
## 2 4.525294 6.675904 5.094062 4.188628 7.075264 4.062682 4.120415 6.678617
## 3 4.945099 6.757165 4.092881 4.239030 4.032549 4.337585 5.024711 4.412778
## 4 5.554609 4.659119 4.229760 8.416807 4.029783 6.542391 5.453046 4.056016
## 5 5.399814 9.364228 4.715699 4.304372 4.721781 4.633272 7.991539 4.713491
```

On rappelle alors la formule de la moyenne empirique :

$$\boxed{\forall n \in \{20, 100, 200\},\ \forall i \in [\![1, B]\!],\ \bar{X}_{n,i} = \frac{1}{n}\sum_{j=1}^{n} X_{j,i}}$$

Ce qui donne, en R :

```
moyEmp <- rep(0,B)

for(i in 1:B) {
  for(j in 1:n) {
    moyEmp[i] = moyEmp[i] + X[j,i]
  }
  moyEmp[i] = moyEmp[i]/n
}

moyEmp_n20 <- moyEmp # On stocke la valeur pour le cas n=20 (histogramme de l'exercice 9)

head(moyEmp) # Affiche les premières valeurs de \bar{X}_{n,i}
```

```
## [1] 4.618423 5.166833 4.890429 5.065637 5.120859 5.187871
```

De même, nous pouvons alors calculer la variance empirique :

$$\boxed{\forall n \in \{20, 100, 200\},\ \forall i \in [\![1, B]\!],\ S_{n,i} = \frac{1}{n}\sum_{j=1}^{n} \left(X_{j,i} - \bar{X}_{n,i}\right)^2}$$

Soit :

```
varEmp <- rep(0,B)

for(i in 1:B) {
  X_barre <- moyEmp[i]
  for(j in 1:n) {
    varEmp[i] <- varEmp[i] + (X[j,i] - X_barre)^2
  }
  varEmp[i] <- varEmp[i]/n
}

head(varEmp) # Affiche les premières valeurs de S_{n,i}
```

```
## [1] 0.4430844 1.6244959 1.5618443 1.5935439 1.2146421 1.3120624
```

Ainsi, nous pouvons simuler cette expérience pour $n = 100$ et $n = 200$.

- Cas $n = 100$

```
# Pareto : 500 échantillons de taille n = 100
n <- 100
X <- data.frame(matrix(nrow = n, ncol = B))

# Stockage des simulations
for(i in 1:B) {
```

```
  X[,i] <- rpareto(n,a,alpha) # n Pareto(a=4,alpha=5)
}

head(X[1:5,1:8]) # Affiche le début des 8 premiers échantillons
```

```
##         X1       X2       X3       X4       X5       X6       X7       X8
## 1 6.994512 5.718522 6.610975 4.497845 4.200599 5.390488 4.434900 4.001805
## 2 4.905261 4.774659 4.585389 5.177081 6.500376 4.280796 4.294253 6.972814
## 3 4.929009 4.030436 4.703301 8.524966 5.513405 7.287330 5.664727 4.051807
## 4 5.438203 5.198094 5.420415 5.083004 4.106917 5.042152 4.339467 4.315486
## 5 5.028952 5.591076 4.170366 7.675764 4.270999 4.352679 4.489886 5.123545
```

```
moyEmp <- rep(0,B)

for(i in 1:B) {
  for(j in 1:n) {
    moyEmp[i] = moyEmp[i] + X[j,i]
  }
  moyEmp[i] = moyEmp[i]/n
}

moyEmp_n100 <- moyEmp # On stocke la valeur pour le cas n=100 (histogramme de l'exercice 9)

head(moyEmp) # Affiche les premières valeurs de \bar{X}_{n,i}
```

```
## [1] 5.134353 4.927831 5.050904 5.011380 4.922763 4.825883
```

```
varEmp <- rep(0,B)

for(i in 1:B) {
  X_barre <- moyEmp[i]
  for(j in 1:n) {
    varEmp[i] <- varEmp[i] + (X[j,i] - X_barre)^2
  }
  varEmp[i] <- varEmp[i]/n
}

head(varEmp) # Affiche les premières valeurs de S_{n,i}
```

```
## [1] 2.737348 1.165566 1.830023 1.532795 2.056044 1.000907
```

- Cas $n = 200$

```
# Pareto : 500 échantillons de taille n = 200
n <- 200
X <- data.frame(matrix(nrow = n, ncol = B))

# Stockage des simulations
for(i in 1:B) {
  X[,i] <- rpareto(n,a,alpha) # n Pareto(a=4,alpha=5)
}

head(X[1:5,1:8]) # Affiche le début des 8 premiers échantillons
```

```
##          X1       X2       X3       X4        X5        X6       X7       X8
## 1 5.169585 4.371174 4.139229 4.295683  8.452253  7.682752 4.448634 4.770845
## 2 4.007433 4.103507 5.503816 6.242967  4.419501 13.996284 4.720248 5.004432
## 3 5.747568 7.215815 4.267508 7.463964  4.505233  4.227089 4.215950 6.590196
## 4 4.049071 9.902913 4.293043 4.291167  4.822446  4.589239 9.336015 4.416269
## 5 4.094907 4.611967 4.392518 4.291233 10.423942  4.011993 6.283167 4.085110
```

```r
moyEmp <- rep(0,B)

for(i in 1:B) {
  for(j in 1:n) {
    moyEmp[i] = moyEmp[i] + X[j,i]
  }
  moyEmp[i] = moyEmp[i]/n
}

moyEmp_n200 <- moyEmp # On stocke la valeur pour le cas n=200 (histogramme de l'exercice 9)

head(moyEmp) # Affiche les premières valeurs de \bar{X}_{n,i}
```

```
## [1] 4.933288 5.000513 5.137627 5.072117 5.118951 5.203624
```

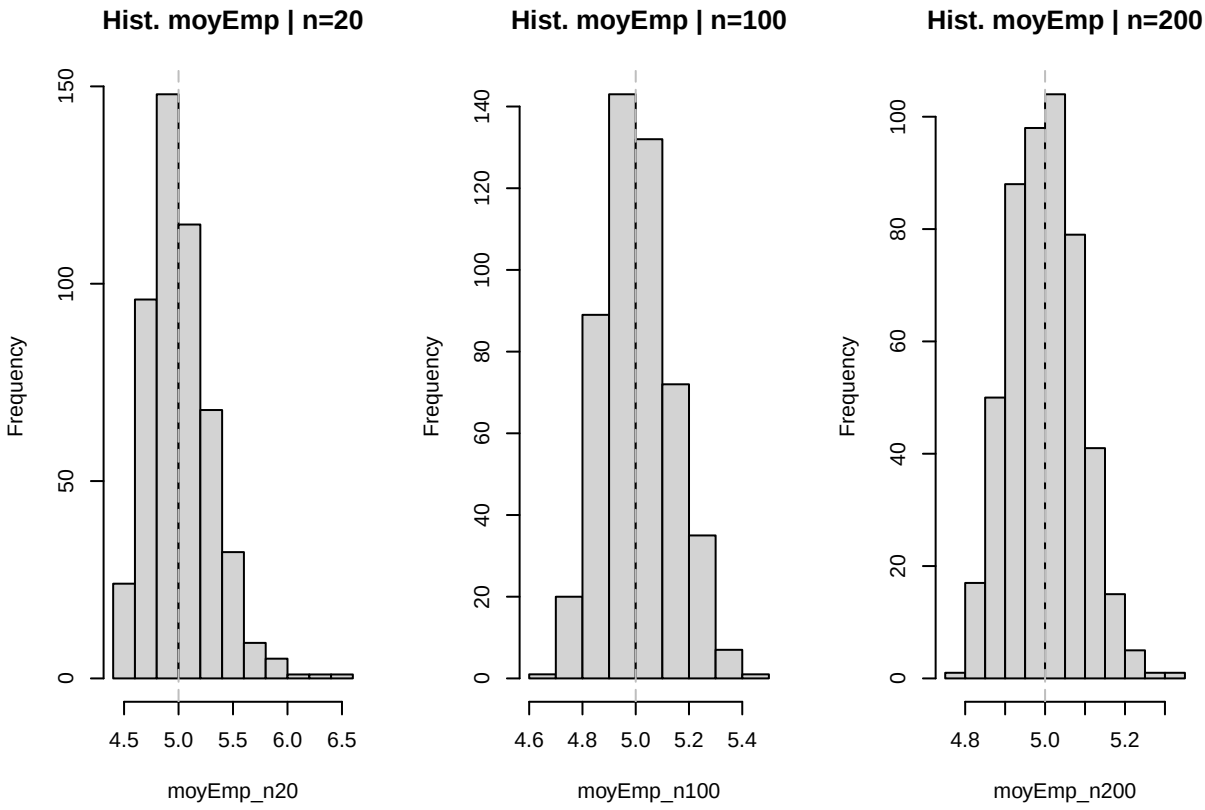
```r
varEmp <- rep(0,B)

for(i in 1:B) {
  X_barre <- moyEmp[i]
  for(j in 1:n) {
    varEmp[i] <- varEmp[i] + (X[j,i] - X_barre)^2
  }
  varEmp[i] <- varEmp[i]/n
}

head(varEmp) # Affiche les premières valeurs de S_{n,i}
```

```
## [1] 1.279393 3.155813 1.738007 2.045270 2.116878 2.364209
```

Exercice 9

On obtient les histogrammes suivants :

```r
par(mfrow=c(1,3)) # 1 x 3 panel

hist(moyEmp_n20, main = "Hist. moyEmp | n=20")
abline(v=5, col="gray", lty=5)
hist(moyEmp_n100, main = "Hist. moyEmp | n=100")
abline(v=5, col="gray", lty=5)
hist(moyEmp_n200, main = "Hist. moyEmp | n=200")
abline(v=5, col="gray", lty=5)
```



Le résultat est cohérent avec l'étude théorique. En effet, avec ces mêmes paramètres, on a :

$$\boxed{\mathbb{E}(X) = \frac{5 \times 4}{4} = 5}$$

Exercice 10

On a supposé que les $X_{n,i}$ sont i.i.d et suivent tous la loi de Pareto(a, α). En supposant qu'on ait $\alpha > 2$, leur espérance et leur variance sont finies.

On pose :

$$\begin{aligned} a_n = \mathbb{E}(X_{n,i}) = \frac{\alpha a}{\alpha - 1} \text{ et } b_n &= \frac{1}{\sqrt{n}} \sqrt{\mathbb{V}(X_{n,i})} \\ &= \frac{1}{\sqrt{n}} \sqrt{\left(\frac{a}{\alpha - 1}\right)^2 \frac{\alpha}{\alpha - 2}} \\ &= \frac{a}{\alpha - 1} \sqrt{\frac{\alpha}{n(\alpha - 2)}} \end{aligned}$$

Soit :

$$\boxed{a_n = \frac{\alpha a}{\alpha - 1}} \text{ et } \boxed{b_n = \frac{a}{\alpha - 1} \sqrt{\frac{\alpha}{n(\alpha - 2)}}}$$

On note $m = \mathbb{E}(X_{n,i}) < +\infty$ et $\sigma^2 = \mathbb{V}(X_{n,i}) < +\infty$. Ainsi, le Théorème Central Limite nous donne :

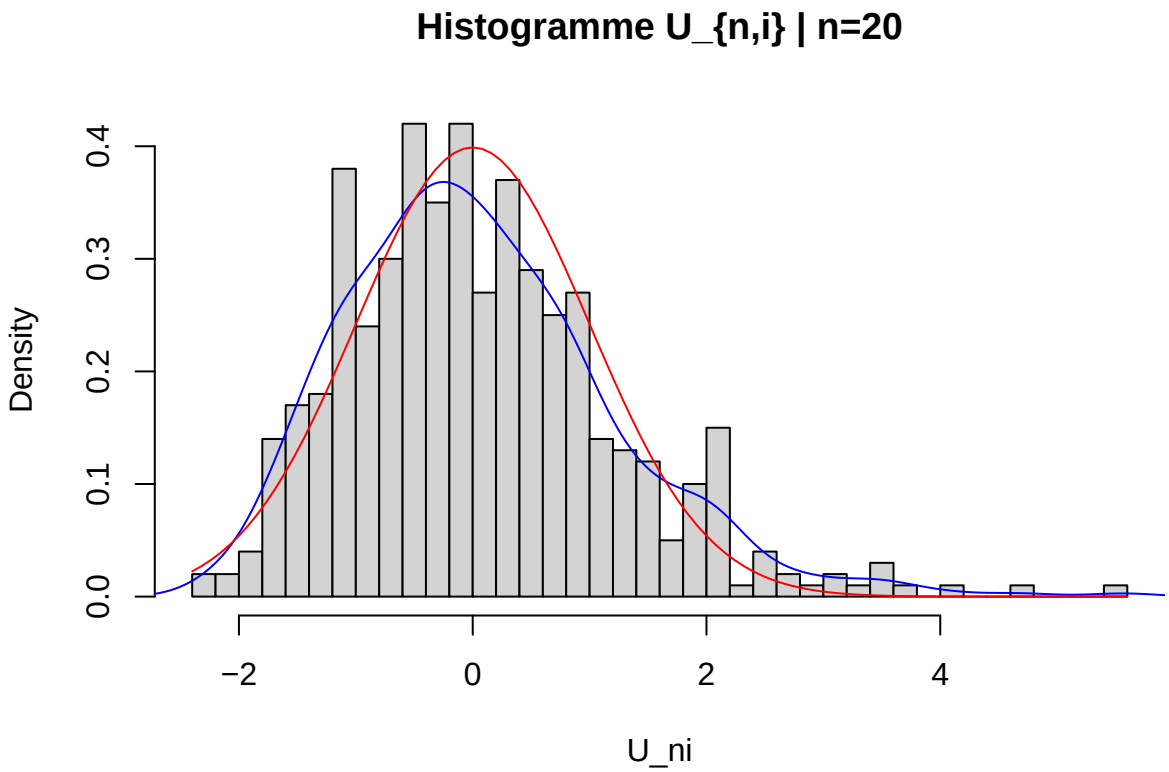
$$U_{n,i} = \frac{\bar{X}_{n,i} - a_n}{b_n} = \sqrt{n}\frac{\bar{X}_{n,i} - m}{\sigma} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

Soit :

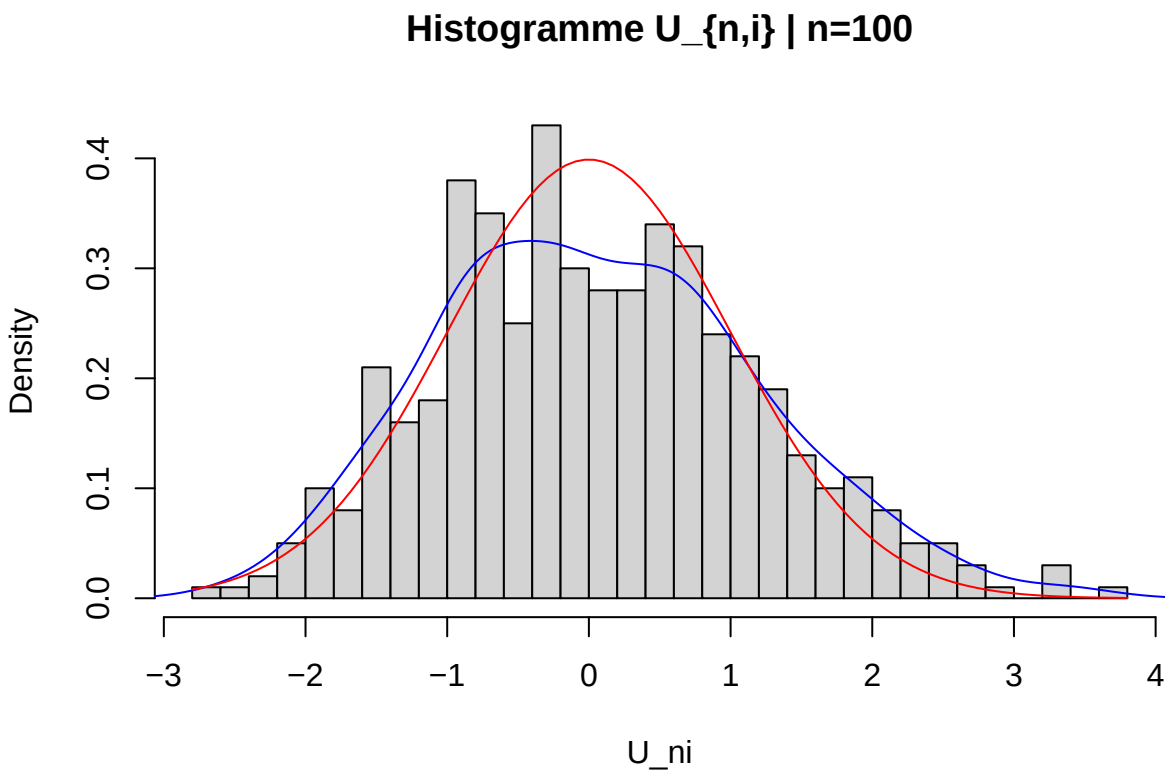
$$\boxed{U_{n,i} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)}$$

Ainsi, on peut tracer les histogrammes des moyennes empiriques normalisées $U_{n,i}$ et la distribution théorique approchée.

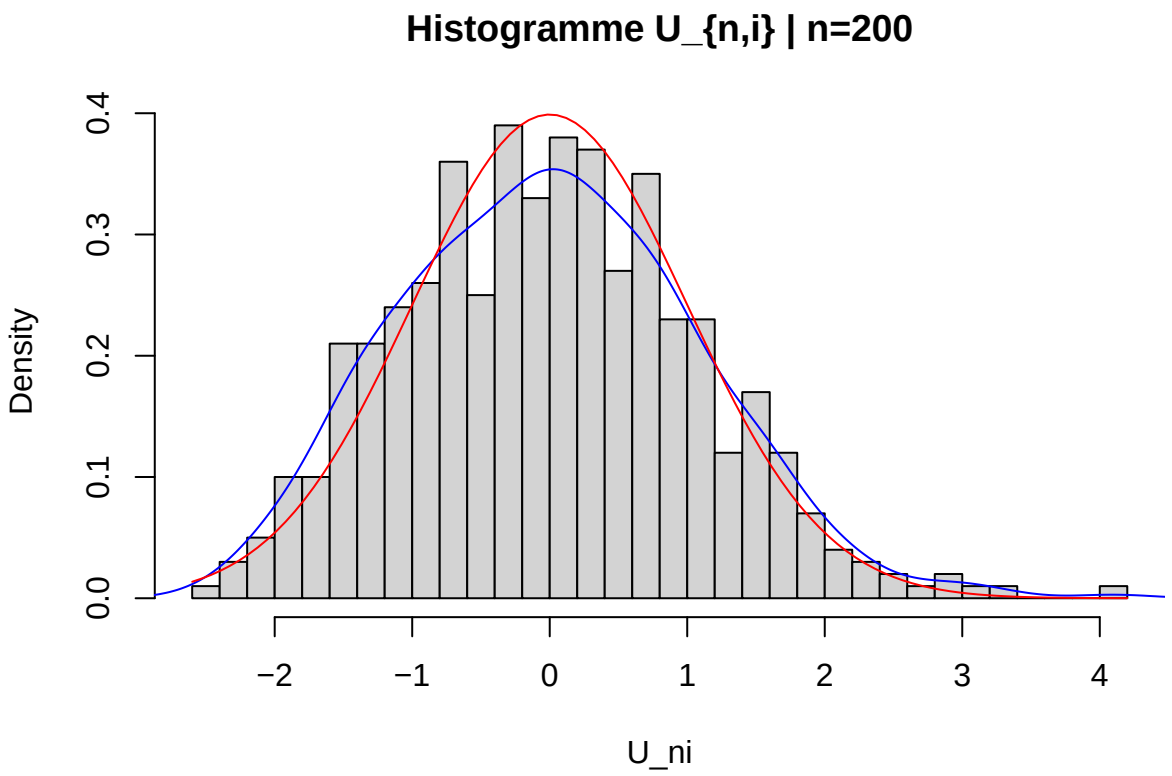
```
n <- 20

a_n <- (alpha*a/(alpha - 1))
b_n <- (a/(alpha - 1))*sqrt(a/(n*(alpha - 2)))
U_ni <- (moyEmp_n20 - a_n)/b_n

h <- hist(U_ni, breaks = 30, main = "Histogramme U_{n,i} | n=20", prob=TRUE)
lines(density(U_ni), col= "blue")
curve(dnorm(x,0,1), add=TRUE, col = "red")
```




```
n <- 100

a_n <- (alpha*a/(alpha - 1))
b_n <- (a/(alpha - 1))*sqrt(a/(n*(alpha - 2)))
U_ni <- (moyEmp_n100 - a_n)/b_n

h <- hist(U_ni, breaks = 30, main = "Histogramme U_{n,i} | n=100", prob=TRUE)
lines(density(U_ni), col= "blue")
curve(dnorm(x,0,1), add=TRUE, col = "red")
```




```r
n <- 200

a_n <- (alpha*a/(alpha - 1))
b_n <- (a/(alpha - 1))*sqrt(a/(n*(alpha - 2)))
U_ni <- (moyEmp_n200 - a_n)/b_n

h <- hist(U_ni, breaks = 30, main = "Histogramme U_{n,i} | n=200", prob=TRUE)
lines(density(U_ni), col= "blue")
curve(dnorm(x,0,1), add=TRUE, col = "red")
```



On constate que pour n suffisamment grand, l'approximation faite tend effectivement vers la loi normale $\mathcal{N}(0,1)$. La qualité de l'approximation est alors bien meilleure.

Quand le Théorème Central Limite ne s'applique pas

Exercice 11

On note n la taille de l'échantillon choisi dans toute cette partie. On choisit dans cette question $n = 50$.

```
theta <- 2 # On fixe theta à 2 dans le cadre de l'exercice
n <- 50
X <- rcauchy(n, theta) # On effectue 50 simulations
moy <- mean(X) # On calcule la moyenne empirique des 50 simulations
moy
```

```
## [1] 1.52671
```

Exercice 12

On fait varier la taille de l'échantillon avec $n = 100$, 1000 et 10000.

```
list_n <- c(100, 1000, 10000)
list_moy <- 1:3
for (i in 1:3) {
  X <- rcauchy(list_n[i], theta)
  list_moy[i] <- mean(X)  # Calcul de la valeur moyenne de la simulation
}
list_moy
```

```
## [1] 1.166867 3.765811 4.052423
```

On remarque alors que les moyennes empiriques obtenues sont diverses et semblent être aléatoires, peu importe la taille n de l'échantillon considéré. La méthode Monte Carlo ne semble alors pas s'appliquer ici. On peut alors émettre l'hypothèse suivante : l'espérance de la loi de Cauchy de paramètre 2 est infinie.

Exercice 13

On sait d'après le cours que l'espérance d'un variable aléatoire X suivant la loi de Cauchy de paramètre 2 est :

$$\mathbb{E}[X] = \frac{\phi_2'(0)}{i}$$

Or la fonction caractéristique d'une loi de Cauchy $\mathcal{C}(\theta)$ s'écrit : $\forall t \in \mathbb{R}$, $\phi_\theta(t) = \exp(i\theta t - |t|)$. A cause du terme $|t|$, cette fonction n'est pas dérivable en 0, et par conséquent, le moment d'ordre 1 de cette probablilité n'existe pas.

Exercice 14

La variable X suivant une loi de Cauchy, la fonction de répartition de X est alors :

$$\forall x \in \mathbb{R},\ F_X(x) = \frac{1}{2} + \frac{1}{\pi} arctan(x - \theta)$$

Alors, la valeur médiane x_M étant caractérisée par : $F_X(x_M) = 0,5$, on a :

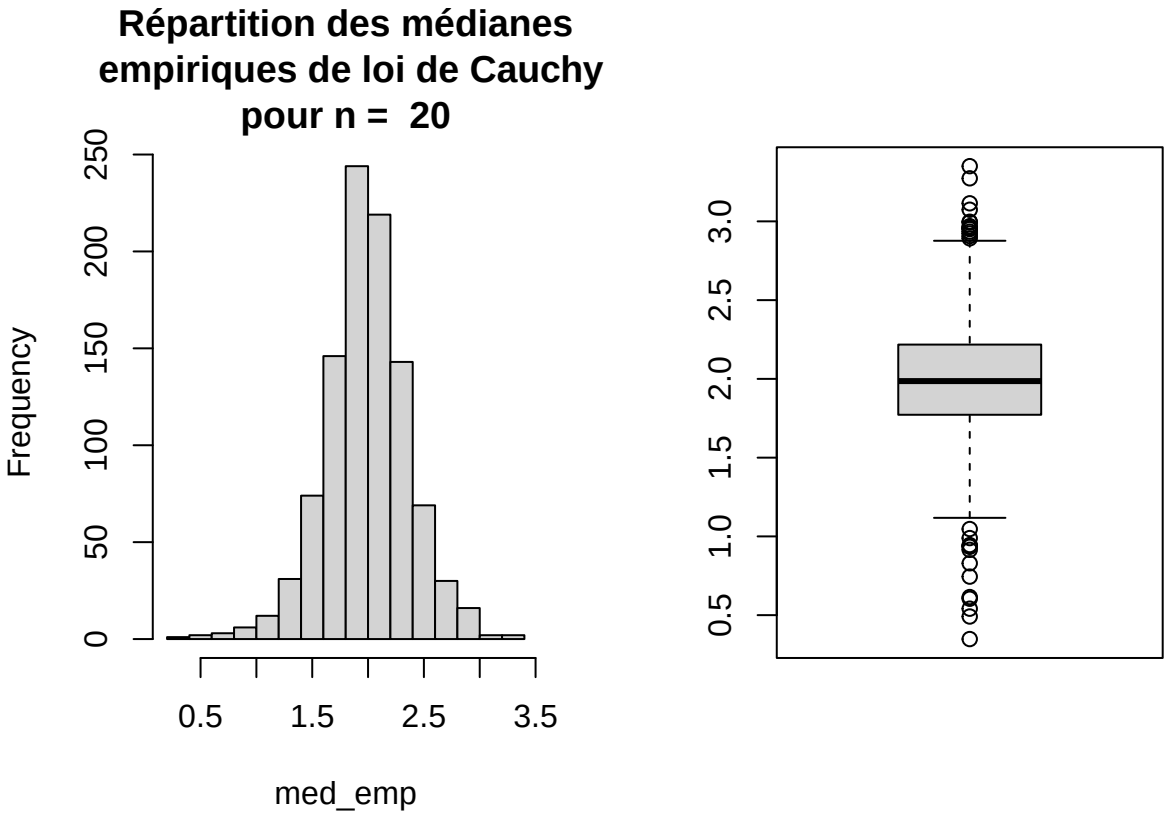
$$\begin{aligned} \frac{1}{2} + \frac{1}{\pi}\arctan(x_M - \theta) = 0,5 &\Leftrightarrow \frac{1}{\pi}\arctan(x_M - \theta) = 0 \\ &\Leftrightarrow arctan(x_M - \theta) = 0 \\ &\Leftrightarrow x_M - \theta = 0 \\ &\Leftrightarrow x_M = \theta \end{aligned}$$

Alors, la médiane d'une loi de Cauchy $\mathcal{C}(\theta)$ vaut θ.

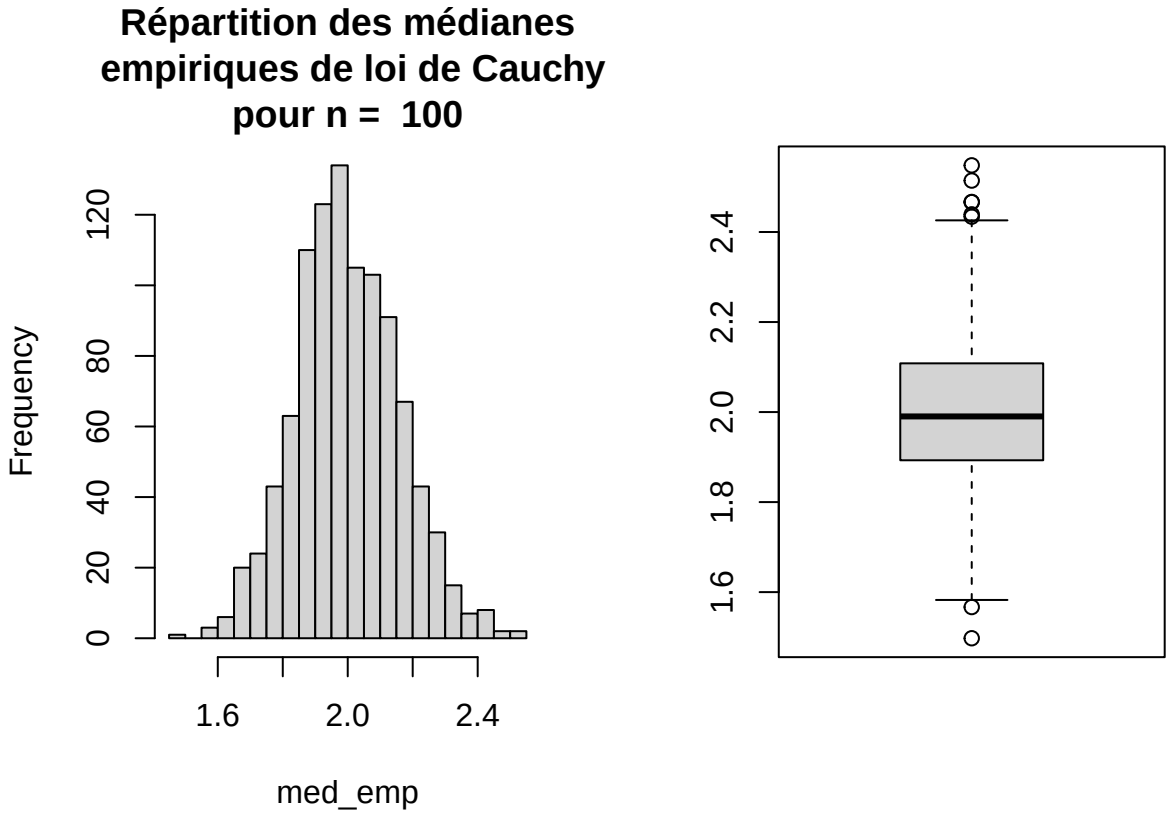
Exercice 15

On considère l'estimateur $\hat{\theta} = \underset{i \in [\![1,n]\!]}{med}(X_i)$, qui calcule empiriquement la médiane des X_i.

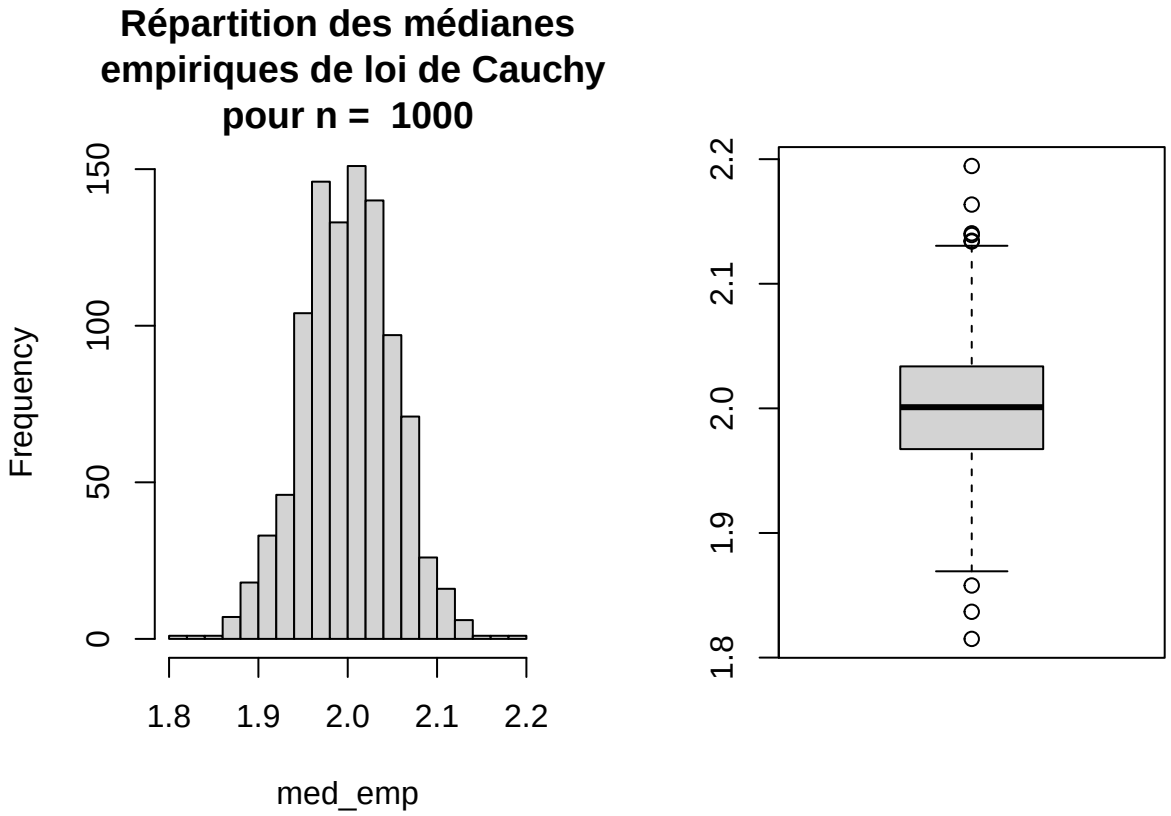
```
N <- 1000 # On réalise N fois la simulation à n valeurs
theta <- 2 # On choisit à nouveau theta = 2
liste_n <- c(20,100,1000)
moy_cauchy <- 1:3
par(mfrow=c(1,2))

for (n in liste_n) {
  med_emp<-1:N
  for (i in 1:N) {
    Xni <- rcauchy(n, theta)
    med_emp[i] <- median(Xni) # Calcul de la valeur médiane de la simulation à n lancers
  }
  hist(med_emp, breaks = 20, main = paste("Répartition des médianes\n empiriques de loi de Cauchy\npour
  boxplot(med_emp) # Affichage des écarts interquartiles
  cat("La moyenne empirique vaut =", med_emp[i], "pour n =", n, ".")
}
```




```
## La moyenne empirique vaut = 2.123329 pour n = 20 .
```




```
## La moyenne empirique vaut = 1.765698 pour n = 100 .
```



```
## La moyenne empirique vaut = 1.929919 pour n = 1000 .
```

Après ces différentes simulations, on constate que la médiane empirique est effectivement proche de θ. On remarque également que plus on réalise de simulations, plus l'écart interquartile diminue. On peut alors considérer cet estimateur comme un bon estimateur.