TP2

Colin Coërchon et Mathis Guillory

2023-02-17

Contents

Simulation et convergence

Variation sous-jacente et échantillonnage répété

Question 1

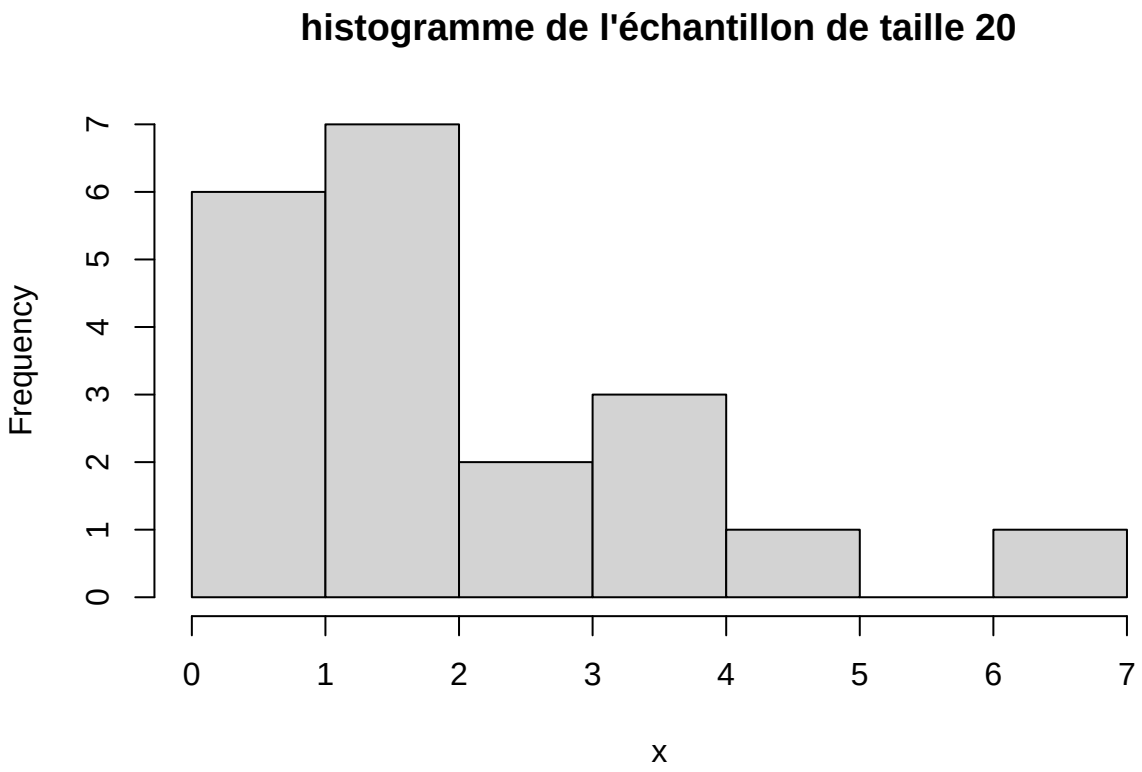
Ici, $\lambda = \frac{1}{2}$, donc on a :

$$\begin{aligned}
\mathbb{P}(X \geq 3) &= \int_3^{+\infty} \frac{1}{2} e^{-\frac{1}{2}x} \mathbb{1}_{\mathbb{R}_+}(x)\, \mathrm{d}x \\
&= 1 - \int_{-\infty}^{3} \frac{1}{2} e^{-\frac{1}{2}x} \mathbb{1}_{\mathbb{R}_+}(x)\, \mathrm{d}x \\
&= 1 - \int_0^3 \frac{1}{2} e^{-\frac{1}{2}x} \mathrm{d}x \\
&= 1 - \frac{1}{2}\left[-2e^{-\frac{1}{2}x}\right]_0^3 \\
\Longrightarrow \mathbb{P}(X \geq 3) &= e^{-\frac{3}{2}}
\end{aligned}$$

La probabilité qu'on observe une valeur supérieure à 3 est égale à $e^{-\frac{3}{2}} \approx 0.22$.

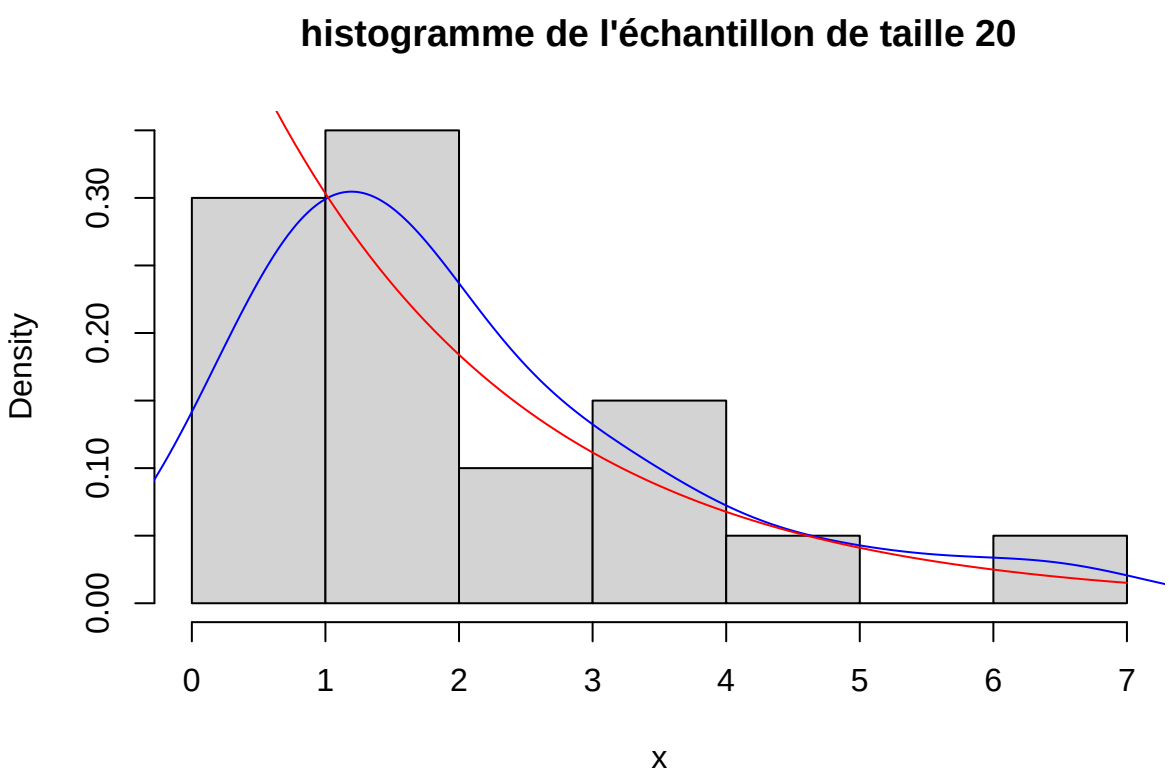
Question 2

```
n <- 20
x <- rexp(n,0.5)
hist(x, main = "histogramme de l'échantillon de taille 20")
```



Superposition des deux :

```
hist(x,main = "histogramme de l'échantillon de taille 20", prob=TRUE)
lines(density(x), col= "blue")
curve(dexp(x,0.5), add=T, col = "red")
```



Probabilité empirique qu'on observe une valeur supérieure à 3 :

```
cat("Voici les valeurs des différents Xi de l'échantillon : ",x)
```

```
## Voici les valeurs des différents Xi de l'échantillon :  1.796766 2.79968 1.255742 3.131418 0.5118806
```

```
Proba <- length(x[x>=3])/n
Proba
```

```
## [1] 0.25
```

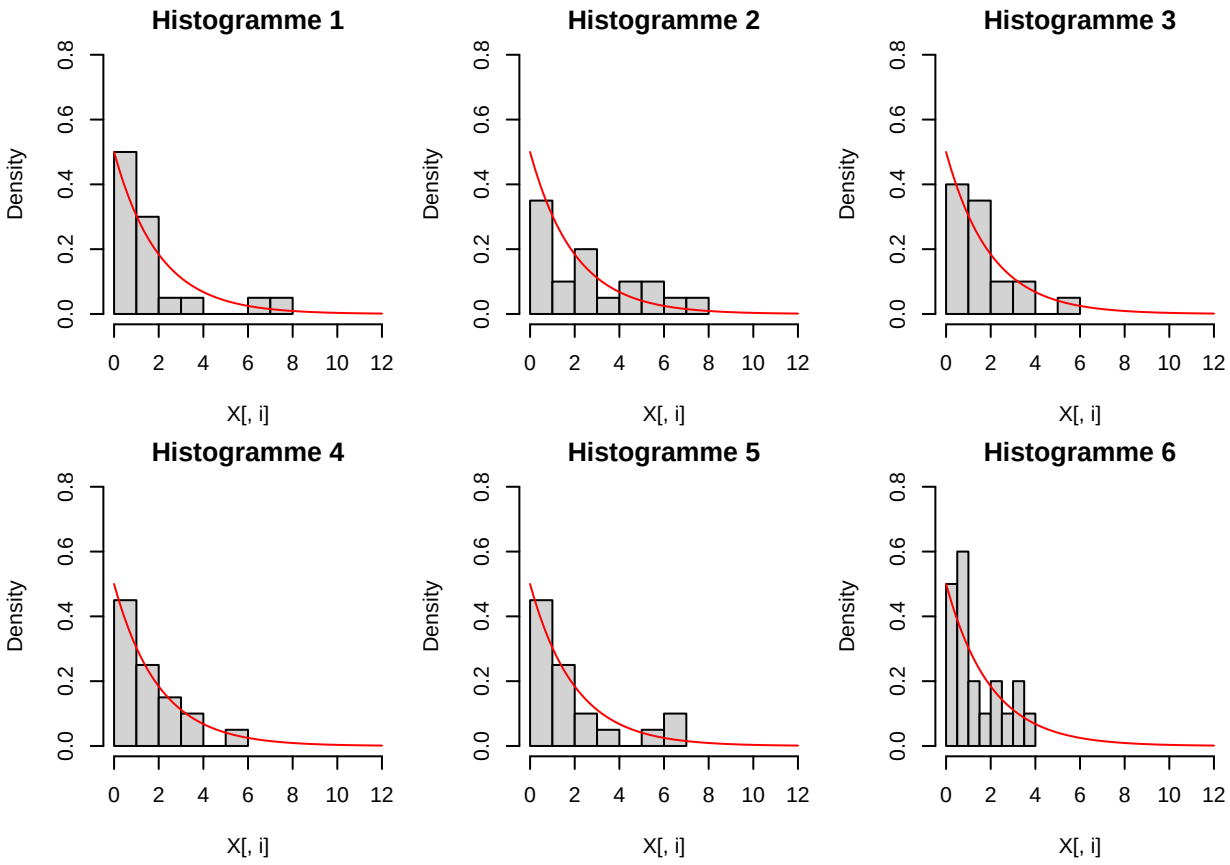
Question 3

La question nous demande de répéter l'opération de la question précédente 5 à 6 fois. Nous avons donc tracé 6 histogrammes différents obtenus par la même opération.

À noter que les axes gardent la même limite pour faciliter la comparaison (les x varient de 0 à 12, et les y varient en densité de 0 à 0.8).

```
nb <- 6
X <- matrix(0,n,nb)
for (i in seq(nb)) {
  X[,i] <- rexp(n,0.5)
}
```

```r
# Configuration de la fenêtre graphique
par(mfrow=c(2,3), mar=c(4,4,2,1))

# Tracé des histogrammes
for(i in 1:6) {
  hist(X[,i],main = sprintf("Histogramme %d",i),xlim=c(0,12),ylim=c(0,0.8),freq=FALSE)
  curve(dexp(x,0.5), add=T, col = "red")
}
```



On peut alors tracer l'histogramme des probabilités empiriques obtenus pour nos 6 essais.

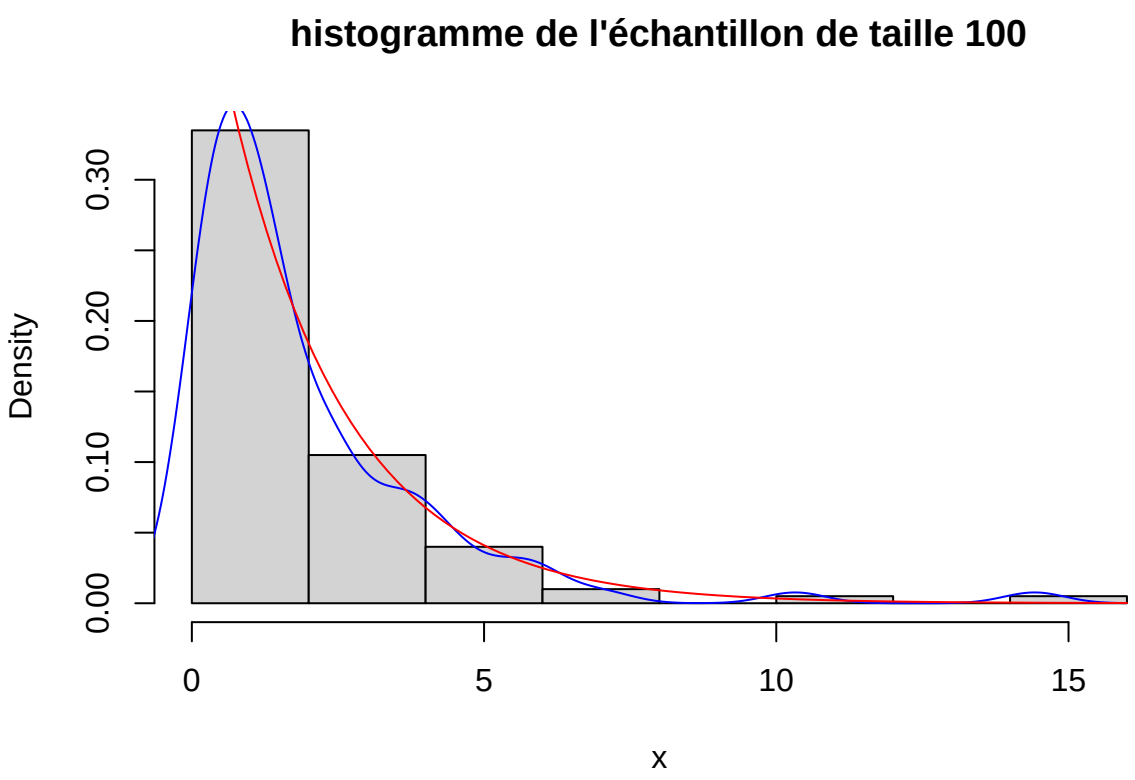
```r
Proba <- rep(0,nb)
for (i in seq(nb)) {
  Xi = X[,i];
  Proba[i] <- length(Xi[Xi>=3])/n
}
Proba
```

```
## [1] 0.15 0.35 0.15 0.15 0.20 0.15
```

```r
hist(Proba,breaks=pretty(Proba),main = "Histogramme des proba empiriques pour les 6 essais")
```

Question 4

```
n <- 100
x <- rexp(n,0.5)

hist(x,main = "histogramme de l'échantillon de taille 100", prob=TRUE)
lines(density(x), col= "blue")
curve(dexp(x,0.5), add=T, col = "red")
```



En comparant les résultats des questions 2 et 3 avec ceux de la question 4, on remarque que, plus la taille de notre échantillon augmente, plus l'histogramme se rapproche davantage de la vraie densité de notre loi exponentielle.

En effet, la courbe de densité de l'histogramme (en bleue) se rapproche davantage de la courbe théorique (en rouge) de notre loi exponentielle lorsque $n = 50$ comparée à la même expérience lorsque $n = 20$.

On peut alors s'intéresser à la probabilité empirique qu'on observe une valeur supérieure à 3 :

```
cat("Voici quelques valeurs des différents Xi de l'échantillon : ",head(x))
```

```
## Voici quelques valeurs des différents Xi de l'échantillon :  1.907883 10.32424 0.7005502 1.148142 0.8
```

```
Proba <- length(x[x>=3])/n
Proba
```

```
## [1] 0.21
```

La valeur de la proba empirique se situe bien davantage aux alentours de 0.22 comparé aux résultat obtenu à la question 2 avec $n = 20$. Cela est rassurant, car $n = 100$ ici, et c'était bien la valeur théorique attendue (calculée à la question 1).

Variabilité aléatoire du maximum de l'échantillon

Question 1

```
x<-runif(10,-1,1)
m<-max(x)
cat("Le maximum de l'échantillon vaut :",m,"\n")
```

```
## Le maximum de l'échantillon vaut : 0.8511597
```

Question 2

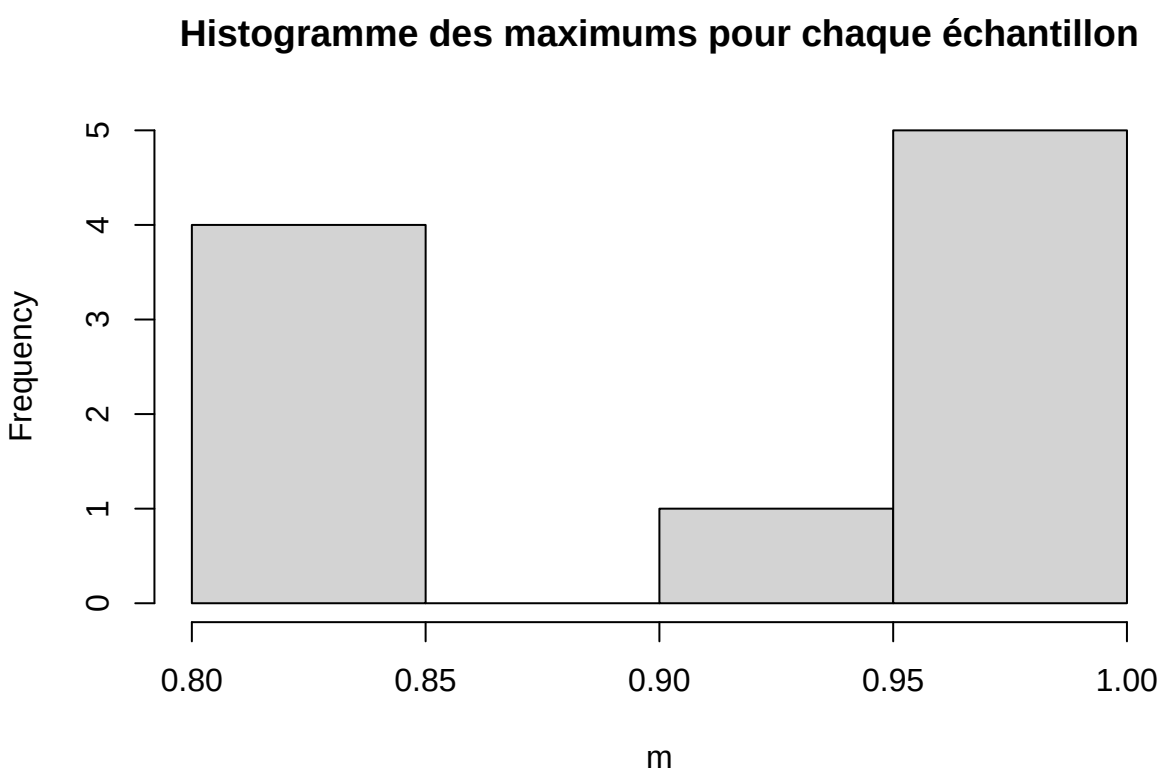
```
n <- 10
x<-seq(n)
res <- matrix(0,10,n)
for (i in x) {
  res[,i] <- runif(10,-1,1)
}
head(res)
```

```
##             [,1]       [,2]       [,3]       [,4]       [,5]       [,6]
## [1,] -0.86835474  0.6557129 -0.7735891  0.1609756 -0.5812027  0.5799061
## [2,] -0.27023144  0.3766984  0.2451777  0.7904707  0.8145283 -0.4119553
## [3,] -0.08458017 -0.4730866 -0.3773122  0.5462545  0.5815381  0.9921662
## [4,] -0.01400042  0.5298403 -0.7495926 -0.4461720  0.5989746  0.2285308
## [5,]  0.65518535 -0.1214041  0.9269898  0.2927060 -0.5047372  0.4712900
## [6,]  0.46304838  0.8481502 -0.4213510  0.8630907  0.4793078 -0.4029317
##              [,7]       [,8]        [,9]       [,10]
## [1,]  0.366851707  0.4770027 -0.79387637 -0.34606353
## [2,] -0.006454989  0.1464102 -0.04384323 -0.69882383
## [3,] -0.081790561 -0.1131603  0.38237053 -0.86951930
## [4,]  0.109589988  0.7364373  0.84045800  0.03112812
## [5,]  0.443311202 -0.1141865  0.96209194 -0.09666804
## [6,]  0.210606338  0.8483839  0.77853639  0.83294559
```

```
m <- rep(0,n)
for (i in x) {
  m[i] <- max(res[,i])
}
cat("Les différentes valeurs des max sont :",m,"\n")
```

```
## Les différentes valeurs des max sont : 0.9182025 0.8481502 0.9589009 0.9814546 0.8145283 0.9921662 0
```

```
hist(m,main="Histogramme des maximums pour chaque échantillon")
```

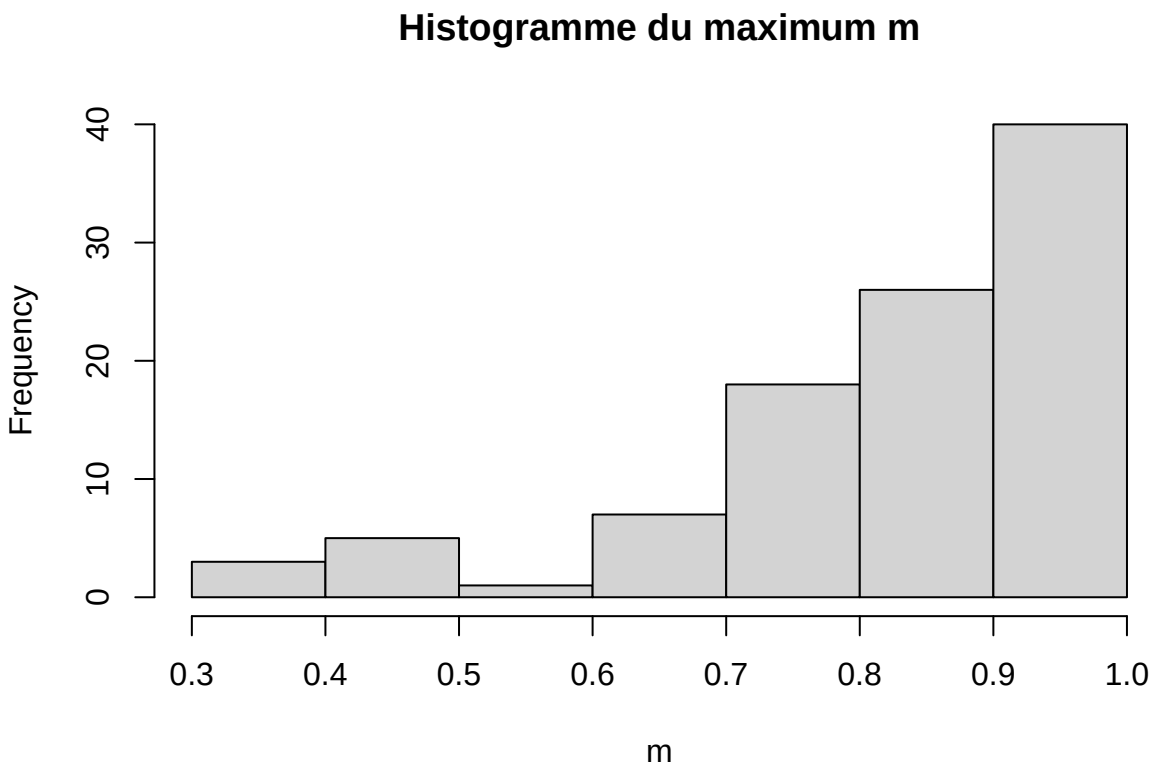


Les résultats des maximums sont plutôt très variables. On s'étale de 0.5 à 1 avec des résultats un peu éparses. Cela s'explique par le fait que, durant les 10 tests, ce sont des échantillons de petites tailles (10) qui ont été analysés. Et 10 tests, c'est aussi plutôt assez limité comme taille d'expérience (d'où les questions 3 et 4 juste après...).

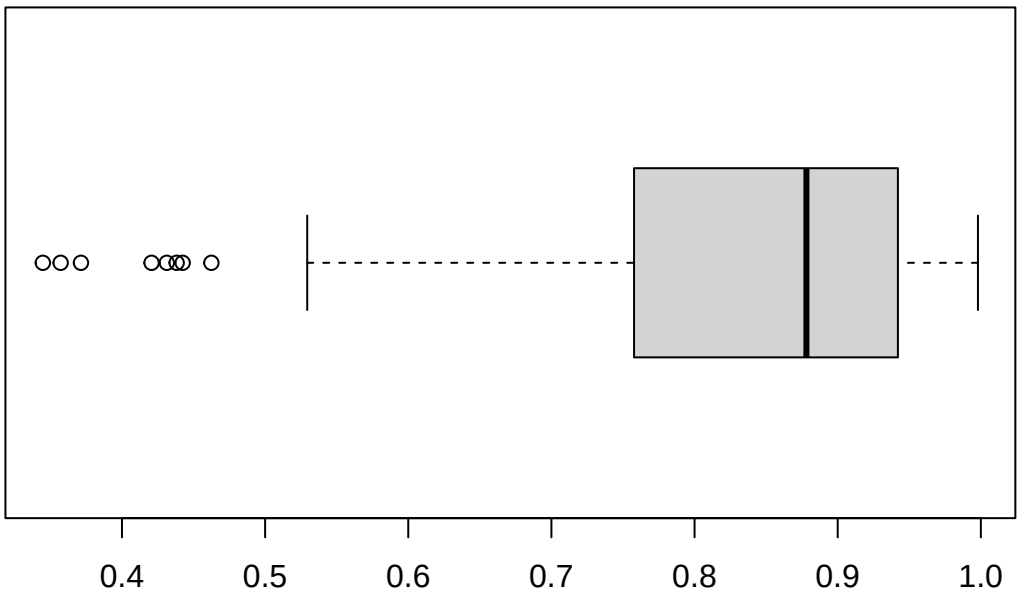
Question 3

On répète non plus 10 fois l'expérience, mais 100 fois maintenant.

```
n <- 100
x<-seq(n)
res <- matrix(0,10,n)
for (i in x) {
  res[,i] <- runif(10,-1,1)
}

m <- rep(0,n)
for (i in x) {
  m[i] <- max(res[,i])
}

hist(m, main = "Histogramme du maximum m")
```




```
boxplot(m, main = "Boîte à moustaches du maximum m", horizontal = T)
```



D'après le TD1, on peut exprimer la densité de M en notant $M = \max_{1 \leq i \leq n} X_i$ avec n le nombre de v.a. uniforme sur $[-1, 1]$ dans notre échantillon.

On sait que pour une seule v.a. $U \sim \mathcal{U}(-1, 1)$, on a :

$$\forall x \in \mathbb{R}, \quad F_U(x) = \mathbb{P}(U \leq x) = \int_{-\infty}^{x} \frac{1}{2} \mathbb{1}_{[-1,1]}(u) \, \mathrm{d}u = \begin{cases} 0 & \text{si } x < -1 \\ \dfrac{x+1}{2} & \text{si } x \in [-1, 1] \\ 1 & \text{si } x > 1 \end{cases}$$

Donc :

$$\forall x \in \mathbb{R}, \quad F_U(x) = \frac{x+1}{2} \mathbb{1}_{[-1,1]}(x) + \mathbb{1}_{]1,+\infty[}(x)$$

On peut alors déterminer la loi du maximum M :

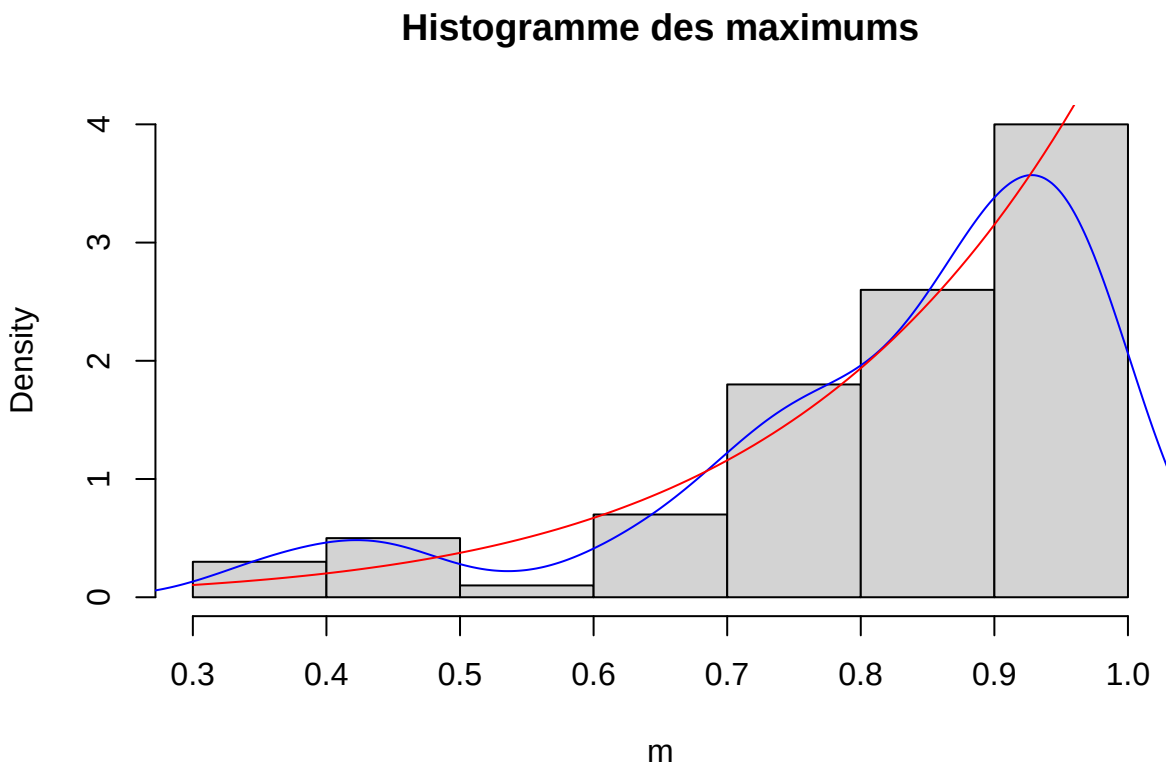
$$
\begin{aligned}
\forall x \in \mathbb{R}, \quad F_M(x) &= \mathbb{P}(M \leq x) \\
&= \mathbb{P}(\max_{1\leq i\leq n} X_i \leq x) \\
&= \mathbb{P}\left(\bigcap_{i=1}^{n} [X_i \leq x]\right) && \text{(par def de l'évènement } \max_{1\leq i\leq n} X_i \leq x) \\
&= \prod_{i=1}^{n} \mathbb{P}(X_i \leq x) && \text{(par indépendance des } X_i) \\
&= \prod_{i=1}^{n} \left(\frac{x+1}{2}\, \mathbb{1}_{[-1,1]}(x) + \mathbb{1}_{]1,+\infty[}(x)\right) && \text{(car les } X_i \text{ sont identiquement distribués)} \\
\implies F_M(x) &= \left(\frac{x+1}{2}\right)^n \mathbb{1}_{[-1,1]}(x) + \mathbb{1}_{]1,+\infty[}(x)
\end{aligned}
$$

Ainsi, on en déduit facilement l'expression de la densité de M:

$$
\forall x \in \mathbb{R}, \quad f_M(x) = \frac{n}{2}\left(\frac{x+1}{2}\right)^{n-1} \mathbb{1}_{[-1,1]}(x)
$$

Superposons alors la densité théorique sur l'histogramme :

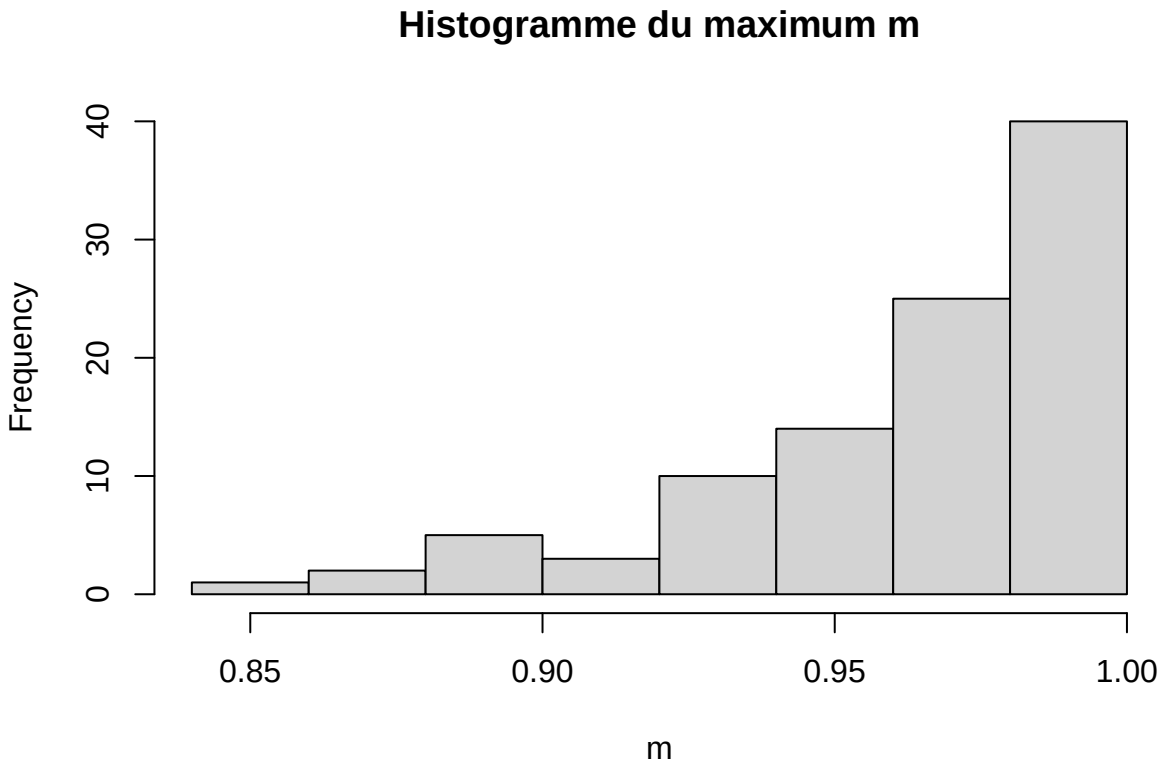
```
n <- 10
unif <- function(x) {n/2 * ((x+1)/2)^(n-1)}
hist(m,main="Histogramme des maximums", prob=T)
lines(density(m),col="blue")
curve(unif(x), add=T, col="red")
```



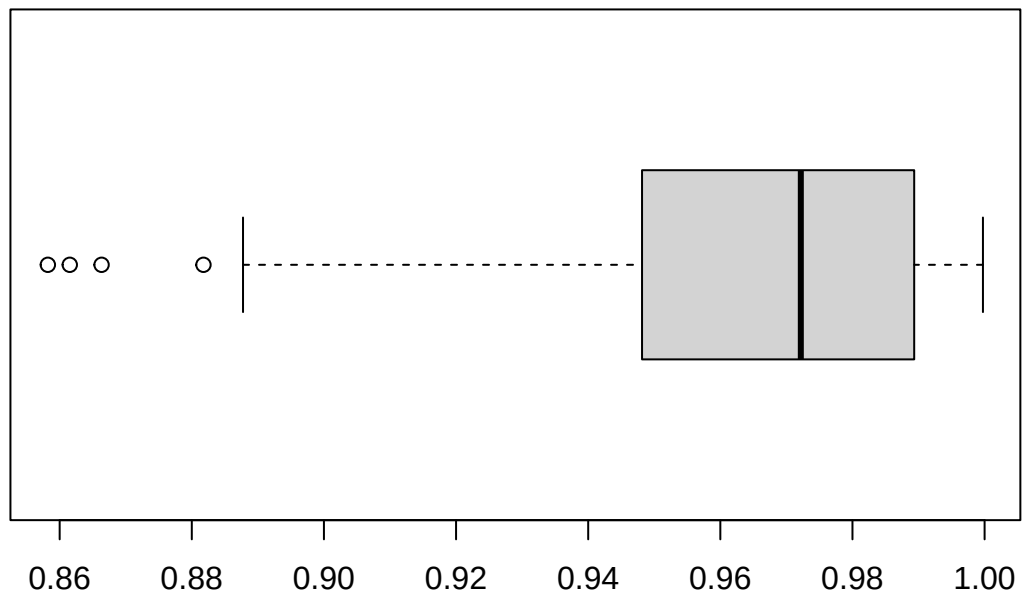
Question 4

On augmente la taille de l'échantillon à 50.

```
n <- 100
x<-seq(n)
res <- matrix(0,50,n)
for (i in x) {
  res[,i] <- runif(50,-1,1)
}

m <- rep(0,n)
for (i in x) {
  m[i] <- max(res[,i])
}

hist(m, main = "Histogramme du maximum m")
```

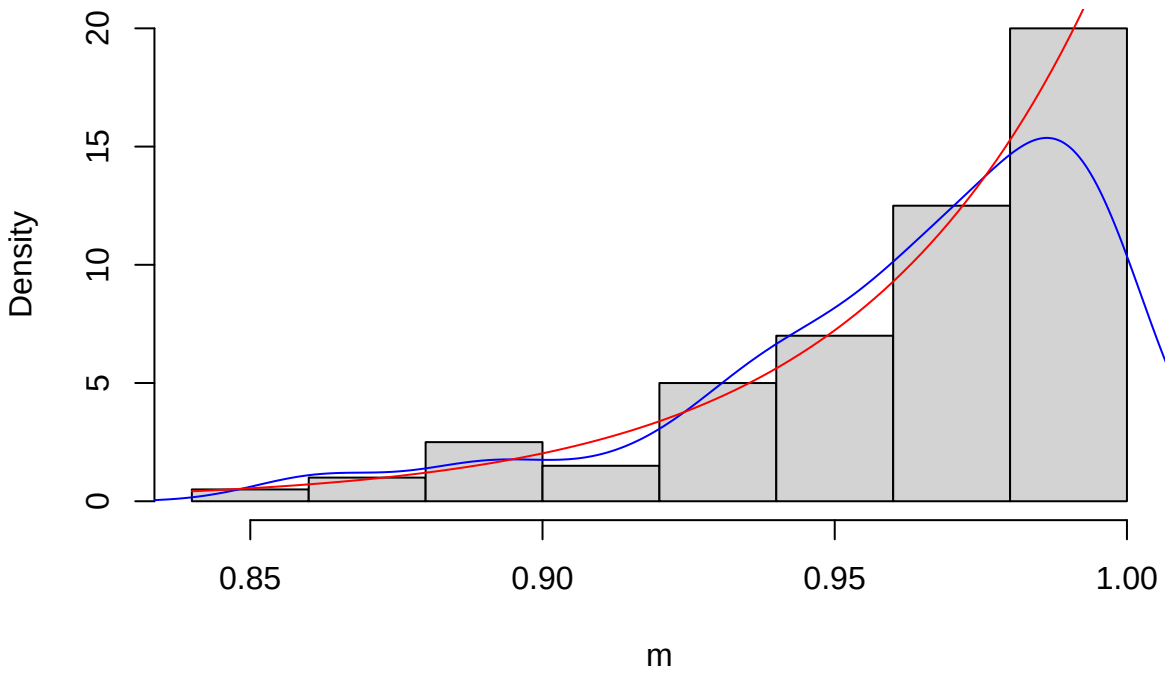



```
boxplot(m, main = "Boîte à moustaches du maximum m",horizontal = T)
```



Et en superposant la densité théorique sur l'histogramme (avec maintenant $n = 50$) :

```
n <- 50
unif <- function(x) {n/2 * ((x+1)/2)^(n-1)}
hist(m, main = "Histogramme du maximum m", prob = TRUE)
lines(density(m),col="blue")
curve(unif(x), add=T, col="red")
```



Comparé à la question 3 où la taille de l'échantillon était $n = 10$, dans la question 4 : $n = 50$. Ainsi, on obtient bien une densité théorique de la fonction max M qui superpose presque parfaitement l'histogramme car on a fixé le nombre de répétitions par X_i à 100, ce qui est assez élevé pour espérer ce genre de résultat.

Cependant, comme $n = 50$, on observe que le maximum M se situe davantage vers la borne supérieure de M qui est 1. En effet, d'après la boîte à moustaches obtenue, plus de 50% des expériences pour calculer M donne une valeur qui se situe au dessus de $0,97$. Alors que la médiane se situe vers $0,88$ lorsque $n = 10$.

Cette différence se voit aussi forcément dans l'expression théorique de notre fonction densité f_M, puisqu'elle est exponentielle en fonction de n.

Monte Carlo Methods

On définit le $\hat{\theta}$ dans la méthode de Monte Carlo comme une moyenne empirique :

$$\theta = \mathbb{E}[\psi(X)] = \int_{\mathbb{R}} \psi(X) f(x)\,\mathrm{d}x \approx \frac{1}{n}\sum_{i=1}^{n} \psi(X_i) = \hat{\theta}$$

On peut alors vérifier que $\mathbb{E}[\hat{\theta}] = \theta$:

$$\begin{aligned}
\mathbb{E}[\hat{\theta}] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\psi(X_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\psi(X_i)] && \text{(par linéarité de l'espérance)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\theta && \text{(par identique distribution)} \\
&= \theta
\end{aligned}$$

Moyenne et phénomène de concentration

Question 1

On suppose que $\sigma^2 = \mathrm{Var}[\psi(X)] < +\infty$. On a alors :

$$\begin{aligned}
\mathbb{P}(|\hat{\theta} - \theta| \geq \delta) &\leq \frac{\mathrm{Var}(\hat{\theta})}{\delta^2} && \text{(Inégalité de Bienaymé-Tchebychev)} \\
&\leq \frac{1}{\delta^2}\,\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n}\psi(X_i)\right] \\
&\leq \frac{1}{n^2\delta^2}\sum_{i=1}^{n}\mathrm{Var}[\psi(X_i)] && \text{(indépendance des } X_i\text{)} \\
&\leq \frac{1}{n^2\delta^2}\sum_{i=1}^{n}\sigma^2 \\
\implies \mathbb{P}(|\hat{\theta} - \theta| \geq \delta) &\leq \frac{\sigma^2}{n\delta^2}
\end{aligned}$$

Question 2

On a :

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\psi(X_i) = \sum_{i=1}^{n}\left(\frac{1}{n}\psi(X_i)\right)$$

Et comme on suppose que : $\forall i \in [\![1, n]\!]$, $a \leq \psi(X_i) \leq b$, on en déduit que :

$$\forall i \in [\![1, n]\!], \quad \frac{a}{n} \leq \frac{1}{n}\psi(X_i) \leq \frac{b}{n}$$

On peut donc appliquer l'inégalité de Hoeffding :

$$\mathbb{P}(|\hat{\theta} - \theta| \geq \delta) \leq 2\exp\left(-\frac{2\delta^2}{\sum_{i=1}^{n}\left(\frac{b-a}{n}\right)^2}\right) = 2\exp\left(-\frac{2n\delta^2}{(b-a)^2}\right)$$

Et donc :

$$\mathbb{P}(|\hat{\theta} - \theta| \geq \delta) \leq 2\exp\left(-\frac{2n\delta^2}{(b-a)^2}\right)$$

Question 3

Pour cette question, on prend $\delta = 2\sigma$. Et on cherche n tel qu'on soit sûr que :

$$\mathbb{P}(|\hat{\theta} - \theta| \geq \delta) \leq \frac{1}{100}$$

Or, par l'inégalité de Bienaymé-Tchebychev utilisée dans la question 1, on sait que :

$$\mathbb{P}(|\hat{\theta} - \theta| \geq \delta) \leq \frac{\sigma^2}{n\delta^2}$$

Donc, en s'assurant que cette borne soit inférieure à 1%, on s'assure l'inégalité $\mathbb{P}(|\hat{\theta} - \theta| \geq \delta) \leq \frac{1}{100}$. On a alors : (avec $\delta = 2\sigma$)

$$\frac{\sigma^2}{n \times (2\sigma)^2} \leq \frac{1}{100} \Longrightarrow \frac{1}{4n} \leq \frac{1}{100} \Longrightarrow 4n \geq 100 \Longrightarrow n \geq 25$$

Ainsi, si $n \geq 25$, on a $\mathbb{P}(|\hat{\theta} - \theta| \geq \delta) \leq \frac{1}{100}$.

Application pour l'estimation de probabilité

Question 1

Pour la partie 1, on avait $X \sim \mathcal{E}(0.5)$, et on s'intéressait à $\mathbb{P}(X \geq 3)$.

Ainsi, on peut écrire :

$$\mathbb{P}(X \geq 3) = \mathbb{E}\left[\mathbb{1}_{\{X \geq 3\}}\right] \qquad \text{(par définition)}$$

Le paramètre d'intérêt θ est donc : $\theta = \mathbb{E}\left[\mathbb{1}_{\{X \geq 3\}}\right]$.

Alors, en écrivant la méthode de Monte Carlo, on a :

$$\begin{aligned} \theta = \mathbb{E}\left[\mathbb{1}_{\{X \geq 3\}}\right] &= \int_{\mathbb{R}} \mathbb{1}_{\{X \geq 3\}} f(x) \mathrm{d}x \\ &\approx \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \geq 3\}} \qquad \text{où } \forall i \in [\![1, n]\!],\ X_i \sim \mathcal{E}(0.5) \end{aligned}$$

On pose donc notre estimateur $\hat{\theta}$ tel que :

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \geq 3\}} \qquad \text{où } \forall i \in [\![1, n]\!],\ X_i \sim \mathcal{E}(0.5)$$

Question 2

On a :

$$\eta = \mathbb{P}(|\hat{\theta} - \theta| \geq \delta) = \mathbb{E}\left[\mathbb{1}_{\left\{|\hat{\theta} - \theta| \geq \delta\right\}}\right] \qquad \text{(par définition)}$$

On pose alors $Z = \mathbb{1}_{\left\{|\hat{\theta} - \theta| \geq \delta\right\}}$. Ainsi, on se retrouve avec :

$$\eta = \mathbb{E}[Z]$$

On utilise alors la méthode de Monte Carlo en écrivant :

$$\eta = \mathbb{E}[Z] \approx \frac{1}{n} \sum_{i=1}^{n} Z_i \qquad \text{où } \forall i \in [\![1, n]\!],\ Z_i = \mathbb{1}_{\left\{|\hat{\theta}_i - \theta| \geq \delta\right\}}.$$

On a exprimé formellement les Z_i en fonction de variables aléatoires que j'ai noté $\hat{\theta}_i$. Pour réaliser notre estimation, il faudrait a priori déterminer "plusieurs estimateurs" $\hat{\theta}_i$.

Or, on peut expliciter ces $\hat{\theta}_i$ en fonction de nos "$X_i \sim \mathcal{E}(0.5)$" de la question précédente pour un peu plus de clarté :

$$\forall i \in [\![1, n]\!], \quad \hat{\theta}_i = \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}_{\{X_{j,i} \geq 3\}}$$

On se retrouve donc avec une matrice de dimension $m \times n$ de variables aléatoires de $\mathcal{E}(0.5)$ indépendantes. On peut prendre arbitrairement $m = n$ pour simplifier un peu (même si prendre m très grand peut donner de meilleurs résultats d'estimation).

On note alors cette matrice $X = (X_{j,i})_{(j,i)\in[\![1,n]\!]^2}$ en ayant pris $m = n$.

Ainsi, dans cette matrice X, on a :

$$\forall (j,i) \in [\![1,n]\!]^2,\ X_{j,i} \sim \mathcal{E}(0.5)$$

Par conséquent, on va pouvoir estimer η par la méthode de Monte Carlo en écrivant :

$$\eta = \mathbb{E}[Z] \approx \frac{1}{n}\sum_{i=1}^{n} Z_i = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{|\hat{\theta}_i - \theta| \geq \delta\}} \qquad \text{où } \forall i \in [\![1,n]\!],\ \hat{\theta}_i = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{X_{j,i}\geq 3\}}$$

De plus, par définition, on a notre paramètre d'intérêt θ qui est identique à la question précédente. Donc : $\theta = \mathbb{P}(X \geq 3) = e^{-\frac{3}{2}}$.

Il ne reste plus qu'à choisir **arbitrairement** un δ pour pouvoir estimer notre probabilité η. En testant, je trouve que prendre $\delta = \frac{1}{n}$ donne des résultats intéressants. Aisni, on prend donc finalement :

$$\delta = \frac{1}{n},\ \theta = e^{-\frac{3}{2}} \text{ et } n \in \{20, 100\}$$

On peut alors déterminer la valeur de η à l'aide de R. Bien que ce η dépende fortement du δ choisi.

Pour $n = 20$:

```
n <- 20
X <- matrix(0,n,n)
theta <- exp(-3/2)
delta <- 1/n

for (i in 1:n) {
  X[,i] <- rexp(n,0.5)
}

theta_hat <- rep(0,n)
# theta_hat correspond au vecteur de taille n comprenant
# les différents theta_hat_i
for (i in 1:n) {
  Xi = X[,i]
  theta_hat[i] <- length(Xi[Xi>=3])/n
}

head(theta_hat)
```

```
## [1] 0.25 0.20 0.35 0.25 0.30 0.30
```

```
eta <- length(theta_hat[abs(theta_hat-theta)>=delta])/n
# On fait ici la moyenne empirique des Z_i,
# ce qui correspond à l'approximation de eta cherchée
cat("L'approximation de eta donne :",eta,"\n")
```

```
## L'approximation de eta donne : 0.5
```

Pour $n = 100$:

```
n <- 100
X <- matrix(0,n,n)
delta <- 1/n

for (i in 1:n) {
  X[,i] <- rexp(n,0.5)
}

theta_hat <- rep(0,n)
# theta_hat correspond au vecteur de taille n comprenant
# les différents theta_hat_i
for (i in 1:n) {
  Xi = X[,i]
  theta_hat[i] <- length(Xi[Xi>=3])/n
}

head(theta_hat)
```

```
## [1] 0.21 0.26 0.25 0.13 0.25 0.25
```

```
eta <- length(theta_hat[abs(theta_hat-theta)>=delta])/n
# On fait ici la moyenne empirique des Z_i,
# ce qui correspond à l'approximation de eta cherchée
cat("L'approximation de eta donne :",eta,"\n")
```

```
## L'approximation de eta donne : 0.75
```

Question 3

En prenant σ qui est l'écart-type d'une variable aléatoire de $\mathcal{E}(0.5)$, on a $\sigma^2 = \dfrac{1}{\left(\frac{1}{2}\right)^2} = 4 \Longrightarrow \sigma = 2$.

Trouvons alors les bornes obtenues avec l'inégalité de Bienaymé-Tchebychev dans notre exemple :

$$\eta = \mathbb{P}(|\hat{\theta} - \theta| \geq \delta) \leq \frac{\sigma^2}{n\delta^2} = \frac{4}{n \times \frac{1}{n^2}} = \frac{4}{\frac{1}{n}}$$
$$\Longrightarrow \eta \leq 4n$$

Prendre $\delta = \dfrac{1}{n}$ a, certes, donné des résultats intéressants pour nos estimations de η. Mais la borne obtenue par l'inégalité de Bienaymé-Tchebychev est bien **respectée**, mais pas très pertinente puisque :

$$\eta \in [0, 1] \quad \text{(puisque } \eta \text{ est une probabilité)} \quad \text{et} \quad \forall n \in \mathbb{N},\ 4n > 1$$

Théorème Central Limite et Estimation Monte Carlo

Question 1

Soit $X \sim \mathcal{P}(a, \alpha)$, de densité $f(x; a, \alpha) = \alpha \frac{a^\alpha}{x^{\alpha+1}} \mathbb{1}_{[a,+\infty[}(x)$.
Premièrement, calculons la fonction de répartition de x :

$$\begin{aligned}
\forall x \in \mathbb{R}, \quad F_X(x) = \mathbb{P}(X \leq x) &= \int_{-\infty}^{x} f(u; a, \alpha)\,\mathrm{d}u \\
&= \int_{-\infty}^{x} \alpha \frac{a^\alpha}{u^{\alpha+1}} \mathbb{1}_{[a,+\infty[}(u)\,\mathrm{d}u \\
&= \alpha a^\alpha \int_a^x \frac{1}{u^{\alpha+1}}\,\mathrm{d}u && \text{(avec } x \geq a\text{)} \\
&= \alpha a^\alpha \left[-\frac{1}{\alpha u^\alpha}\right]_a^x \\
&= \alpha a^\alpha \left(-\frac{1}{\alpha x^\alpha} + \frac{1}{\alpha a^\alpha}\right) \\
&= 1 - \left(\frac{a}{x}\right)^\alpha
\end{aligned}$$

La fonction de répartition de la loi de Pareto $P(a, \alpha)$ est donc donnée par :

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

On suppose que $\alpha > 1$. Calculons alors l'espérance de X.

$$\begin{aligned}
\mathbb{E}[X] &= \int_{\mathbb{R}} x f(x; a, \alpha)\,\mathrm{d}x && \text{(par définition)} \\
&= \int_0^{+\infty} \mathbb{P}(X > t)\,\mathrm{d}t && \text{(propriété de l'espérance, car } X \text{ est positive)} \\
&= \int_0^{+\infty} 1 - \mathbb{P}(X \leq t)\,\mathrm{d}t \\
&= \int_0^{+\infty} 1 - \left(1 - \left(\frac{a}{t}\right)^\alpha\right) \mathbb{1}_{[a,+\infty[}(t)\,\mathrm{d}t \\
&= \int_0^a 1\,\mathrm{d}t + \int_a^{+\infty} \left(\frac{a}{t}\right)^\alpha \mathrm{d}t \\
&= a + a^\alpha \left[\frac{t^{1-\alpha}}{1-\alpha}\right]_a^{+\infty} \\
&= a + \frac{a}{\alpha - 1} && \text{(car } \alpha > 1\text{)} \\
&= \frac{(\alpha-1)a + a}{\alpha - 1} \\
\implies \mathbb{E}[X] &= \frac{\alpha a}{\alpha - 1}
\end{aligned}$$

Bonus :

Attention ! Il est noté dans l'énoncé que $\mathrm{Var}(X) = \left(\frac{\alpha a}{\alpha - 1}\right)^2 \frac{\alpha}{\alpha - 2}$. Mais il y a en réalité une petite coquille (un α en trop).

Recalculons donc la variance de X. Et commençons par calculer $\mathbb{E}[X^2]$.

On suppose que $\alpha > 2$, on a :

$$\begin{aligned}
\mathbb{E}[X^2] = \int_{\mathbb{R}} x^2 f(x; a, \alpha)\,\mathrm{d}x &= \int_{\mathbb{R}} x^2 \alpha \frac{a^\alpha}{x^{\alpha+1}} \mathbb{1}_{[a,+\infty[}(x)\,\mathrm{d}x && \text{(par définition)} \\
&= \alpha a^\alpha \int_a^{+\infty} x^{1-\alpha}\,\mathrm{d}x \\
&= \alpha a^\alpha \left[\frac{x^{2-\alpha}}{2-\alpha}\right]_a^{+\infty} \\
&= \alpha a^\alpha \frac{a^{2-\alpha}}{\alpha - 2} && \text{(car } \alpha > 2\text{)} \\
&= \frac{\alpha a^2}{\alpha - 2}
\end{aligned}$$

Ainsi, on a :

$$\begin{aligned}
\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 &= \frac{\alpha a^2}{\alpha - 2} - \frac{\alpha^2 a^2}{(\alpha - 1)^2} && \text{(par définition)} \\
&= \left(\frac{a}{\alpha - 1}\right)^2 \left(\frac{\alpha(\alpha - 1)^2}{\alpha - 2} - \alpha^2\right)
\end{aligned}$$

Et, on a :

$$\frac{\alpha(\alpha - 1)^2}{\alpha - 2} - \alpha^2 = \frac{\alpha^3 - 2\alpha^2 + \alpha - \alpha^2(\alpha - 2)}{\alpha - 2} = \frac{\alpha}{\alpha - 2}$$

Et ainsi, on en déduit la *vraie* variance de X :

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \left(\frac{a}{\alpha - 1}\right)^2 \frac{\alpha}{\alpha - 2}$$

Question 2

On simule $N = 1000$ échantillons de v.a.r. i.i.d. de loi commune Pareto $\mathcal{P}(a, \alpha)$. Ces échantillons seront alors de taille n avec $n \in \{5, 30, 100\}$.

Attention ! Il est nécessaire d'installer un *package* pour pouvoir utiliser la commande `rpareto`. Nous avons choisi le *package* 'EnvStats'.

De plus, le choix des paramètres a et α est laissé libre. Il peut être pertinent de prendre $\alpha > 2$ pour que nos moyennes et variances empiriques aient plus de sens.

En effet, d'après la question 1, si $X \sim \mathcal{P}(a, \alpha)$, alors X admet une espérance si et seulement si $\alpha > 1$, et X admet une variance si et seulement si $\alpha > 2$.

On prendra donc arbitrairement :

$$a = 4 \quad \text{et} \quad \alpha = 3$$

On commence par $n = 5$:

```r
#Pareto: 1000 échantillons de taille n = 5
N <- 1000
n <- 5
X <- matrix(0, n, N)

for(i in 1:N) {
 X[ ,i] <- rpareto(n,4,3)
}

cat("Aperçu de nos 10 premiers échantillons :\n")
```

```
## Aperçu de nos 10 premiers échantillons :
```

```r
head(X[,1:10])
```

```
##           [,1]     [,2]     [,3]     [,4]      [,5]     [,6]     [,7]     [,8]
## [1,]  9.371536 5.752144 7.387686 4.819170  5.961181 4.018560 5.772415 4.823324
## [2,]  9.859486 9.524295 4.889792 6.702554  5.147362 7.608171 4.368440 8.739271
## [3,]  5.094723 4.150851 4.129393 5.827295  6.116952 9.924694 5.492156 5.836251
## [4,]  8.117113 5.109872 5.891042 5.858321 17.052638 4.264220 4.620015 7.770799
## [5,] 15.541826 8.047903 4.730857 4.486316  6.516510 4.265364 5.026258 5.057136
##          [,9]    [,10]
## [1,] 4.237712 5.326140
## [2,] 5.786180 4.292836
## [3,] 4.154364 9.690456
## [4,] 4.293591 6.241797
## [5,] 6.173939 6.671650
```

On a créé une grande matrice $(X_{j,i})_{(j,i)}$ de taille $n \times N$ dans laquelle chaque colonne i est un échantillon de taille n. Ainsi, on peut alors exprimer la moyenne empirique $\overline{X_{n,i}}$ pour chaque échantillon i :

$$\forall n \in \mathbb{N}^*, \forall i \in [\![1, N]\!], \quad \overline{X_{n,i}} = \frac{1}{n} \sum_{j=1}^{n} X_{j,i}$$

Ce qui se traduit en R par :

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

for(i in 1:N) {
  for(j in 1:n) {
    moyEmp[i] <- moyEmp[i] + X[j,i]
  }
  moyEmp[i] <- moyEmp[i]/n
}

# On peut plus simplement écrire :
# for(i in 1:N) { moyEmp[i] <- mean(X[ ,i]) }

moyEmp1 <- moyEmp #Pour l'histogramme après...

cat("Aperçu des moyennes empiriques des premiers échantillons :\n")
```

```
## Aperçu des moyennes empiriques des premiers échantillons :
```

```
head(moyEmp)
```

```
## [1] 9.596937 6.517013 5.405754 5.538731 8.158929 6.016202
```

De la même manière, on peut exprimer la variance empirique $S_{n,i}$ de chaque échantillon i :

$$\forall n \in \mathbb{N}^*, \forall i \in [\![1, N]\!], \quad S_{n,i} = \frac{1}{n} \sum_{j=1}^{n} \left(X_{j,i} - \overline{X_{n,i}}\right)^2$$

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

for(i in 1:N) {
  xBarre <- mean(X[ ,i])
  for(j in 1:n) {
    varEmp[i] <- varEmp[i] + (X[j,i] - xBarre)^2
  }
  varEmp[i] <- varEmp[i]/n
}

# On peut plus simplement écrire :
# for(i in 1:N) { moyEmp[i] <- var(X[ ,i]) }

cat("Aperçu des variances empiriques des premiers échantillons :\n")
```

```
## Aperçu des variances empiriques des premiers échantillons :
```

```
head(varEmp)
```

```
## [1] 11.5842514  3.9102327  1.3028722  0.6330472 19.9729818  5.5872235
```

Pour $n = 30$:

```
#Pareto: 1000 échantillons de taille n = 30
N <- 1000
n <- 30
X <- matrix(0, n, N)

for(i in 1:N) {
 X[ ,i] <- rpareto(n,4,3)
}

cat("Aperçu de nos 10 premiers échantillons :\n")
```

```
## Aperçu de nos 10 premiers échantillons :
```

```r
head(X[,1:10])
```

```
##          [,1]      [,2]     [,3]     [,4]      [,5]      [,6]      [,7]
## [1,] 5.020227  4.292736 7.098773 4.944433  4.723084  5.362609 26.014045
## [2,] 5.402094  4.168106 4.049350 6.766204 11.197799  5.918516  7.566928
## [3,] 4.039682  5.794529 5.040024 6.821148  5.877379  8.381182  4.011116
## [4,] 5.123608  5.887661 5.474308 5.967302  4.148066 10.041891  7.574710
## [5,] 4.039673  4.611491 4.552031 6.282164  7.024307  7.228189  4.727700
## [6,] 4.600864 15.883628 4.266848 4.533829  4.255848  6.045231  4.651635
##           [,8]     [,9]    [,10]
## [1,]  4.697693 4.155491 4.110875
## [2,]  4.458002 4.027985 4.149044
## [3,]  7.294456 5.854476 6.022317
## [4,]  4.524029 4.556368 6.813624
## [5,] 31.997968 4.118584 8.586988
## [6,]  9.979351 8.257433 5.987130
```

```r
#La moyenne empirique
moyEmp <- rep(0,N)

for(i in 1:N) {
  for(j in 1:n) {
    moyEmp[i] <- moyEmp[i] + X[j,i]
  }
  moyEmp[i] <- moyEmp[i]/n
}

moyEmp2 <- moyEmp #Pour l'histogramme après...

cat("Aperçu des moyennes empiriques des premiers échantillons :\n")
```

```
## Aperçu des moyennes empiriques des premiers échantillons :
```

```r
head(moyEmp)
```

```
## [1] 6.747382 6.500789 5.892794 5.626590 5.628078 6.633390
```

```r
#La variance empirique
varEmp <- rep(0,N)

for(i in 1:N) {
  xBarre <- mean(X[ ,i])
  for(j in 1:n) {
    varEmp[i] <- varEmp[i] + (X[j,i] - xBarre)^2
  }
  varEmp[i] <- varEmp[i]/n
}

cat("Aperçu des variances empiriques des premiers échantillons :\n")
```

```
## Aperçu des variances empiriques des premiers échantillons :
```

```r
head(varEmp)
```

```
## [1] 42.502650  9.637131  6.067630  4.847862  3.586732  5.372371
```

Pour $n = 100$:

```r
#Pareto: 1000 échantillons de taille n = 100
N <- 1000
n <- 100
X <- matrix(0, n, N)

for(i in 1:N) {
 X[ ,i] <- rpareto(n,4,3)
}

cat("Aperçu de nos 10 premiers échantillons :\n")
```

```
## Aperçu de nos 10 premiers échantillons :
```

```r
head(X[,1:10])
```

```
##           [,1]      [,2]     [,3]     [,4]     [,5]     [,6]     [,7]     [,8]
## [1,]  5.179714  4.495182 5.607880 4.521097 4.437385 4.477970 8.705251 6.301924
## [2,]  4.825752  4.263370 6.075616 6.260656 7.678125 5.001758 8.733515 7.099508
## [3,] 13.020550  4.814187 6.949337 4.849253 6.309910 6.432228 4.897235 7.449618
## [4,]  5.420706  4.162542 4.768424 4.492192 4.132037 5.224729 5.661595 4.213178
## [5,]  4.211505 31.107441 4.029204 7.187698 4.041641 4.015232 4.414975 4.006598
## [6,]  5.272252  4.693861 4.337294 7.125514 7.059010 5.236051 4.490388 5.739214
##          [,9]    [,10]
## [1,] 4.832995 4.014683
## [2,] 6.098976 4.916003
## [3,] 4.055208 4.872154
## [4,] 4.615006 4.522248
## [5,] 4.244382 4.072413
## [6,] 5.338195 4.072894
```

```r
#La moyenne empirique
moyEmp <- rep(0,N)

for(i in 1:N) {
  for(j in 1:n) {
    moyEmp[i] <- moyEmp[i] + X[j,i]
  }
  moyEmp[i] <- moyEmp[i]/n
}

moyEmp3 <- moyEmp #Pour l'histogramme après...

cat("Aperçu des moyennes empiriques des premiers échantillons :\n")
```

```
## Aperçu des moyennes empiriques des premiers échantillons :
```

```
head(moyEmp)
```

```
## [1] 6.692657 7.596411 5.988288 6.000510 5.734061 5.307629
```

```
#La variance empirique
varEmp <- rep(0,N)

for(i in 1:N) {
  xBarre <- mean(X[ ,i])
  for(j in 1:n) {
    varEmp[i] <- varEmp[i] + (X[j,i] - xBarre)^2
  }
  varEmp[i] <- varEmp[i]/n
}

varEmp3 <- varEmp

cat("Aperçu des variances empiriques des premiers échantillons :\n")
```

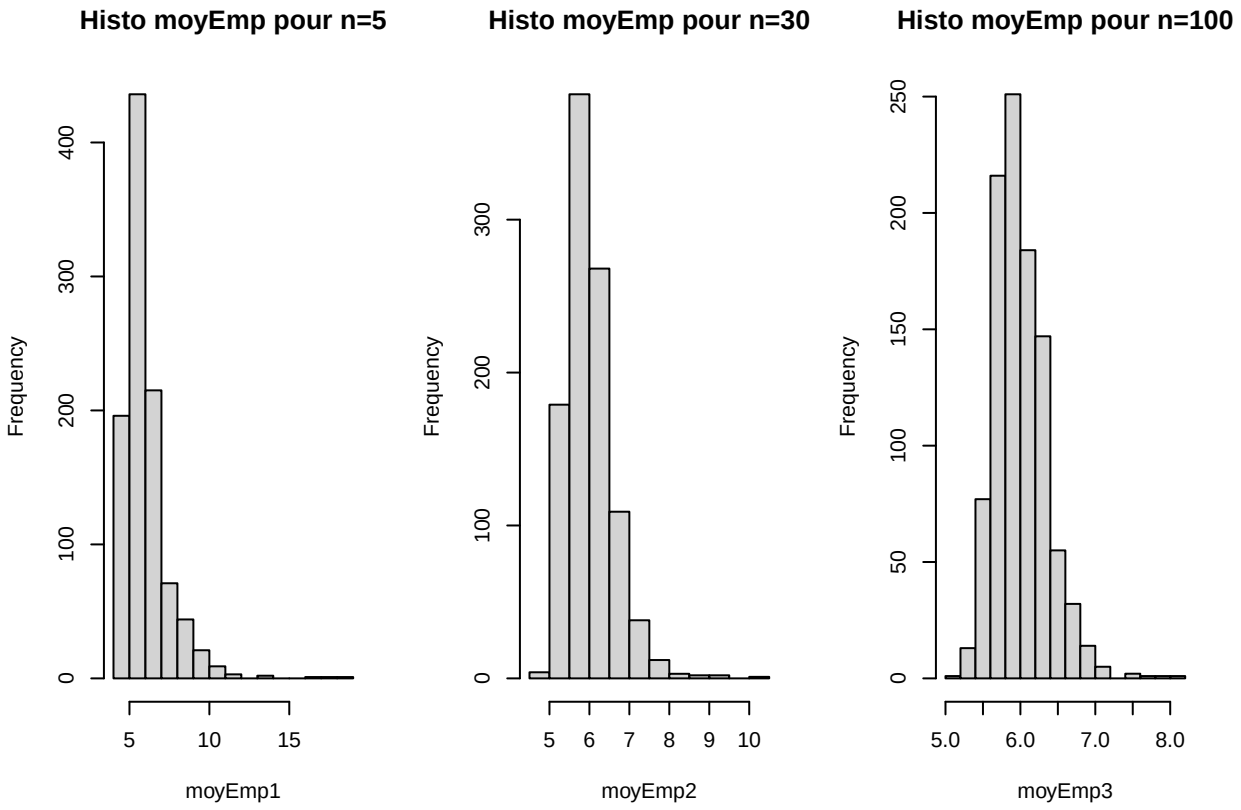
```
## Aperçu des variances empiriques des premiers échantillons :
```

```
head(varEmp)
```

```
## [1] 14.843153 80.809607  8.231443  6.989311  4.134312  1.582425
```

Question 3

```
# Configuration de la fenêtre graphique
par(mfrow=c(1,3))

# Tracé des histogrammes
hist(moyEmp1, main = "Histo moyEmp pour n=5")
hist(moyEmp2, main = "Histo moyEmp pour n=30")
hist(moyEmp3, main = "Histo moyEmp pour n=100")
```



Plus n grandit, plus on a des valeurs qui se rapproche de 6. Ce qui est tout à fait logique, car, si $X \sim \mathcal{P}(a, \alpha)$ avec $a = 4$ et $\alpha = 3$, on a :

$$\mathbb{E}[X] = \frac{\alpha a}{\alpha - 1} = \frac{3 \times 4}{2} = 6$$

De même, pour la variance, on remarque que la moyenne des variances empiriques qu'on trouve pour nos 1000 échantillons se rapproche de la valeur 12 lorsque n grandit. En effet, pour $n = 100$, on a :

```
cat("Moyenne des variances des 1000 échantillons lorsque n = 100 :",mean(varEmp3),"\n")
```

```
## Moyenne des variances des 1000 échantillons lorsque n = 100 : 11.13791
```

Ce qui est tout à fait logique, car, si $X \sim \mathcal{P}(a, \alpha)$ avec $a = 4$ et $\alpha = 3$, on a :

$$\mathrm{Var}(X) = \left(\frac{a}{\alpha - 1}\right)^2 \frac{\alpha}{\alpha - 2} = \left(\frac{4}{2}\right)^2 \frac{3}{3 - 2} = 12$$

Question 4

On cherche $\forall n \in \mathbb{N}$ a_n et b_n positifs tels que $U_{n,i}$ converge vers une certaine loi. Avec :

$$\forall N \in \mathbb{N}, \forall i \in [\![1, N]\!], \quad U_{n,i} = \frac{\overline{X_{n,i}} - a_n}{b_n}$$

Or, on sait que lorsque $X_1, X_2, \ldots, X_n$ sont des v.a.r. i.i.d., et que $\forall i \in [\![1, n]\!]$, $\mathbb{E}(X_i) = m < +\infty$ et $\text{Var}(X_i) = \sigma^2 < +\infty$, le Théorème Central Limite nous donne :

$$\sqrt{n}\,\frac{\overline{X_n} - m}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Et ici, les $X_{n,i}$ sont i.i.d. et suivent tous la même loi de Pareto $\mathcal{P}(a, \alpha)$. L'espérance et la variance des $X_{n,i}$ sont donc finies. Il reste donc à correctement exprimer a_n et b_n pour pouvoir appliquer le TCL.

On a : [en prenant la variance correcte]

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

En posant donc $\forall n \in \mathbb{N}$, $a_n = \dfrac{\alpha a}{\alpha - 1}$ et $b_n = \dfrac{a}{\alpha - 1}\sqrt{\dfrac{\alpha}{n(\alpha - 2)}}$, on a :

$$U_{n,i} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Voici le code R pour $n = 5$:

(C'est le même pour $n = 30$ et $n = 100$ en changeant simplement la valeur de n.)

```
N <- 1000
n <- 5
a <- 4
alpha <- 3
X <- matrix(0, n, N)

for(i in 1:N) {
 X[ ,i] <- rpareto(n,a,alpha)
}

Xbar <- rep(0,N)

for(i in 1:N) {
  Xbar[i] <- mean(X[ ,i])
}

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

U_ni <- (Xbar - a_n)/b_n

cat("Aperçu des premiers U_ni :\n")
```

```
## Aperçu des premiers U_ni :
```

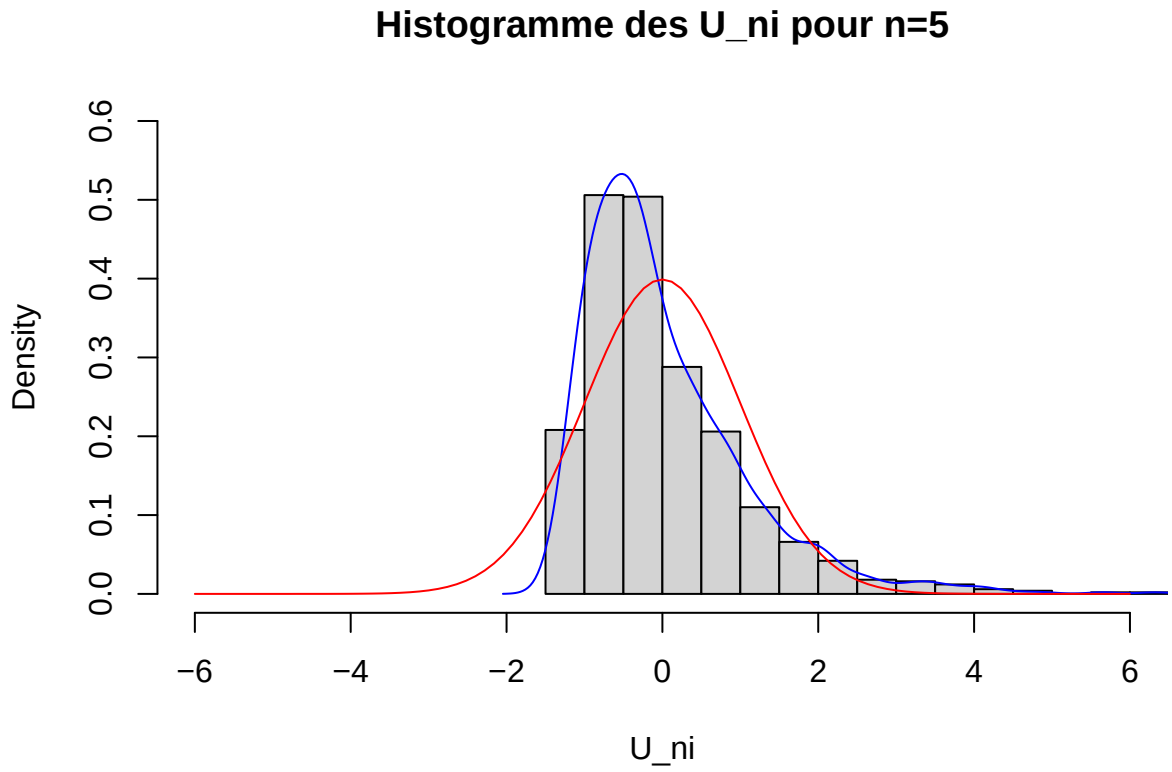
```
head(U_ni)
```

```
## [1]  0.07743601  1.40213859 -1.22105797  0.40625629 -0.05762415  0.96395221
```

Ainsi, $U_{n,i}$ est un vecteur de taille N qui correspond à notre échantillon renormalisé. On trace alors l'histogramme des $U_{n,i}$ et on superpose cet histogramme avec la loi normale $\mathcal{N}(0,1)$ pour analyser la vitesse de convergence du Théorème Central Limite en fonction de n.

Pour $n = 5$:

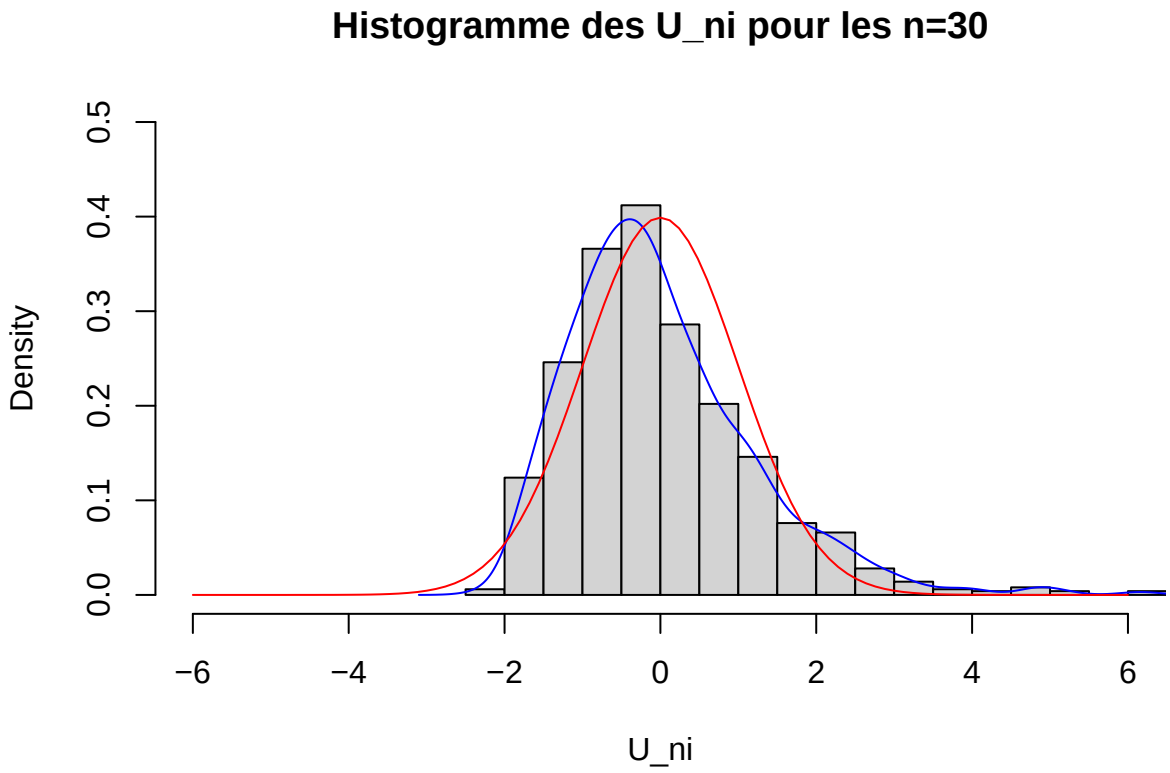
```
hist(U_ni, breaks = 30, main = "Histogramme des U_ni pour n=5", xlim = c(-6,6), ylim = c(0,0.6), prob=T)
lines(density(U_ni), col= "blue")
curve(dnorm(x,0,1), add=T, col = "red")
```



Pour $n = 30$:

```
n <- 30

X <- matrix(0, n, N)

for(i in 1:N) {
 X[ ,i] <- rpareto(n,a,alpha)
}

Xbar <- rep(0,N)
```

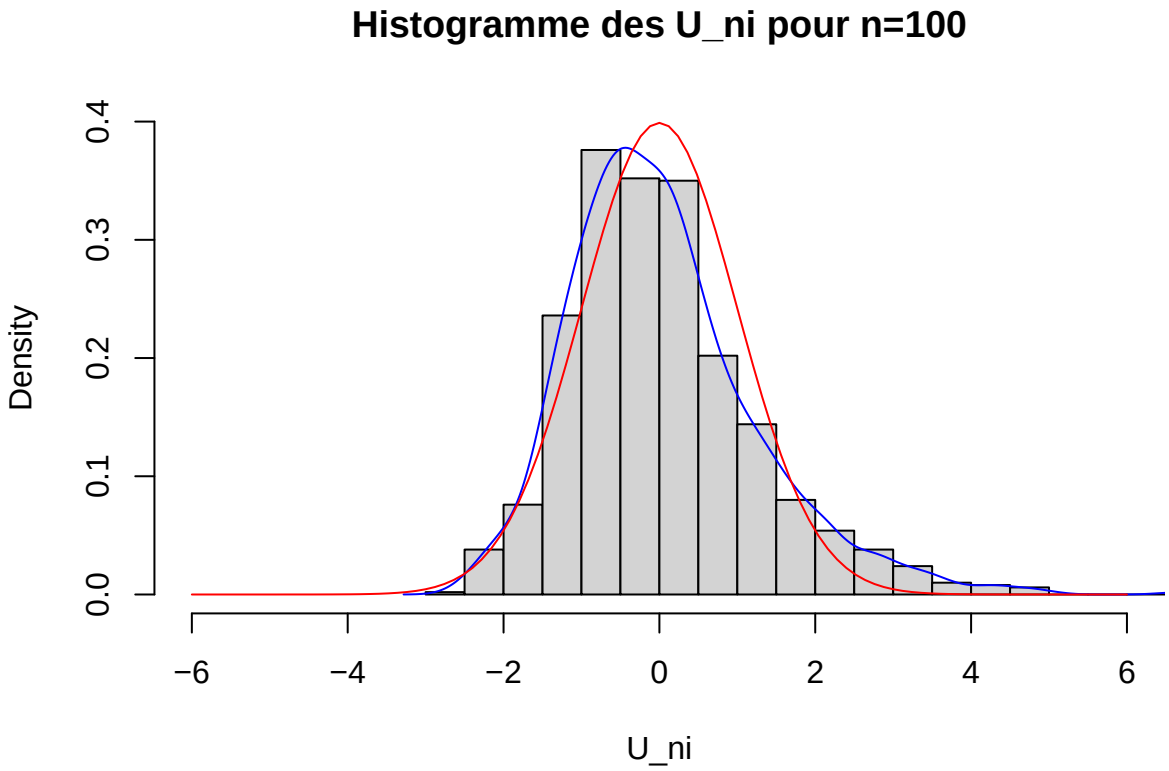
```r
for(i in 1:N) {
  Xbar[i] <- mean(X[ ,i])
}

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

U_ni <- (Xbar - a_n)/b_n

hist(U_ni, breaks=30, main = "Histogramme des U_ni pour les n=30", xlim = c(-6,6), ylim = c(0,0.5), prob
lines(density(U_ni), col= "blue")
curve(dnorm(x,0,1), add=T, col = "red")
```



Pour $n = 100$:

```r
n <- 100

X <- matrix(0, n, N)

for(i in 1:N) {
 X[ ,i] <- rpareto(n,a,alpha)
}

Xbar <- rep(0,N)
```

```r
for(i in 1:N) {
  Xbar[i] <- mean(X[ ,i])
}

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

U_ni <- (Xbar - a_n)/b_n

hist(U_ni, breaks=30, main = "Histogramme des U_ni pour n=100", xlim = c(-6,6), ylim = c(0,0.4), prob=T)
lines(density(U_ni), col= "blue")
curve(dnorm(x,0,1), add=T, col = "red")
```



Ainsi, on remarque que n a une grande influence sur la qualité de l'approximation des $U_{n,i}$ par la loi normale $\mathcal{N}(0,1)$.

Pour $n = 5$, l'approximation est plutôt mauvaise, tandis que pour $n = 30$ et encore mieux pour $n = 100$, la courbe expérimentale bleue tend à s'approcher de la courbe théorique rouge de la loi centrée réduite lorsque n est grand. La convergence en loi donnée par le TCL s'opère effectivement dans cet exemple pour n grand.

Quand le théorème de central limite ne s'applique pas

On s'intéresse à la loi de Cauchy $\mathcal{C}(\theta)$ de densité : $\forall x \in \mathbb{R}$, $f(x,\theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}$.

Question 1

```
n <- 20
theta <- 2

x <- rcauchy(n,theta)

cat("Aperçu de notre échantillon pour n = 20 :\n")
```

```
## Aperçu de notre échantillon pour n = 20 :
```

```
head(x)
```

```
## [1]  1.7544911  2.0181969  1.5460559 -0.5340029  0.2785628 22.1596091
```

Calculons alors la moyenne empirique de notre échantillon :

```
moyEmp <- mean(x)
cat("Moyenne empirique de l'échantillon :",moyEmp,"\n")
```

```
## Moyenne empirique de l'échantillon : 2.0224
```

Question 2

```
tab_moyEmp <- rep(0, 4)
tab_n <- list(20, 100, 1000, 10000)
cpt <- 1

for(n in tab_n) {
  x <- rcauchy(n, theta, 1)
  tab_moyEmp[cpt] <- mean(x)
  cpt <- cpt + 1
}

cat("Tableau des 4 moyennes théoriques :\n")
```

```
## Tableau des 4 moyennes théoriques :
```

```
tab_moyEmp
```

```
## [1] 1.283916 2.681398 9.952097 1.451149
```

On remarque que les moyennes empiriques prennent des valeurs un peu éparses, avec l'impression que les moyennes empiriques ne semblent pas converger vers une certaine valeur lorsque n est grand.

En quelque sorte, il semble que la méthode de Monte Carlo pour estimer $\mathbb{E}[X]$ lorsque $X \sim \mathcal{C}(\theta)$ ne fonctionne pas.

Question 3

On sait que la fonction caractéristique de la loi de Cauchy $\mathcal{C}(\theta)$ est donnée par :

$$\varphi_X(t) = \mathbb{E}\left[\mathrm{e}^{\mathrm{i}Xt}\right] = \int_{-\infty}^{+\infty} f(x,\theta)\,\mathrm{e}^{\mathrm{i}xt}\,\mathrm{d}x = \mathrm{e}^{\mathrm{i}\theta t - |t|}$$

Et, on sait que pour toute variable aléatoire X :

Si X admet une espérance,
alors sa fonction caractéristique φ_X est dérivable en 0, avec $\mathbb{E}[X] = -\varphi'_X(0)$.

Et, ici, on a $\varphi_X(t) = \exp(\mathrm{i}\theta t - |t|) = \exp(\mathrm{i}\theta t)\exp(-|t|)$.

Or, la fonction $t \mapsto |t|$ n'étant pas dérivable en 0, la fonction $t \mapsto \exp(-|t|)$ n'est aussi pas dérivable en 0 par composition. Donc, on en déduit que la fonction φ_X n'est pas dérivable en 0.

Par contraposition, on en déduit que X **n'admet pas d'espérance**.

Ainsi, le fait que **la loi de Cauchy n'admette pas d'espérance** explique les résultats très éparses que nous avons obtenu durant la question précédente.

Question 4

Pour déterminer la médiane d'une loi de probabilité, une bonne méthode peut être de résoudre l'équation $F_X(m) = \frac{1}{2}$ avec m la médiane cherchée, et F_X la fonction caractéristique de X.

Premièrement, déterminons la fonction caractéristique de $X \sim \mathcal{C}(\theta)$:

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\pi}\frac{1}{1+(t-\theta)^2}\,\mathrm{d}t = \frac{1}{\pi}\left[\arctan(t-\theta)\right]_{-\infty}^{x} = \frac{1}{\pi}\arctan(x-\theta) + \frac{1}{2}$$

On peut ainsi déterminer la médiane de X :

$$\begin{aligned} F_X(m) = \frac{1}{2} &\iff \frac{1}{\pi}\arctan(x-\theta) + \frac{1}{2} = \frac{1}{2} \\ &\iff \arctan(x-\theta) = 0 \\ &\iff m - \theta = 0 \qquad \text{(car arctan est bijective de } \mathbb{R} \text{ dans }]-\tfrac{\pi}{2}, \tfrac{\pi}{2}[\,) \\ &\iff m = \theta \end{aligned}$$

Donc la médiane d'une loi de Cauchy $\mathcal{C}(\theta)$ est θ.

Question 5

On cherche un estimateur $\hat{\theta}$ de θ lorsqu'on se trouve avec une variable aléatoire X qui suit une loi de Cauchy de paramètre θ.

Or, la question précédente nous informe que la médiane de la loi de Cauchy $\mathcal{C}(\theta)$ est θ. Ainsi, on va poser :

$$\hat{\theta} = \text{"}\textit{la médiane de } X\text{"} = F_X^{-1}\left(\tfrac{1}{2}\right)$$

En sachant que, pour échantillon de taille n, l'estimateur de la médiane est donné par le $\frac{n+1}{2}$-ème élément lorsqu'on trie les observations par ordre croissant. On peut donc l'exprimer comme :

$$m = X_{(n+1)/2}$$

On prendra arbitrairement $\theta = 2$.

Pour $n = 20$:

```
n <- 20
theta <- 2

X <- rcauchy(n,theta)
m <- median(X)

cat("La médiane de notre échantillon pour n = 20 est :",m,"\n")
```

```
## La médiane de notre échantillon pour n = 20 est : 2.449874
```

Pour $n = 100$:

```
n <- 100
theta <- 2

X <- rcauchy(n,theta)
m <- median(X)

cat("La médiane de notre échantillon pour n = 20 est :",m,"\n")
```

```
## La médiane de notre échantillon pour n = 20 est : 2.00457
```

Pour $n = 1000$:

```
n <- 1000
theta <- 2

X <- rcauchy(n,theta)
m <- median(X)

cat("La médiane de notre échantillon pour n = 20 est :",m,"\n")
```

```
## La médiane de notre échantillon pour n = 20 est : 1.957569
```

On remarque que, plus n est grand, plus on se rapproche du paramètre d'intérêt θ (même s'il peut encore y avoir des valeurs un peu dispersées en fonction des tests).

Pour évaluer la qualité de notre estimateur statistique, on peut utiliser l'EQM (Erreur Quadratique Moyenne).

On la définit de cette manière :

$$\text{EQM}(\hat{\theta}) := \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$$

Et on va alors déterminer cette erreur d'estimation pour chacune des valeurs de n en prenant arbitrairement $N = 1000$ répétitions par échantillon.

```
n <- c(20, 100, 1000)
theta <- 2
N <- 1000

med <- matrix(0, N, length(n))
eqmr <- matrix(0, N, length(n))

for (i in seq_along(n)) {
  for (j in 1:N) {
    x <- rcauchy(n[i], theta)
    med[j,i] <- median(x)
    eqmr[j,i] <- ((med[j,i] - theta)/theta)^2
  }
}

# On prend la moyenne de chaque colonne à l'aide de apply()
# pour obtenir les résultats finaux
med_mean <- apply(med, 2, mean)
eqmr_mean <- apply(eqmr, 2, mean) * 100

data.frame(n, med_mean, eqmr_mean)
```

```
##      n med_mean  eqmr_mean
## 1   20 2.023474 3.16225234
## 2  100 2.000142 0.63047737
## 3 1000 2.000871 0.06687844
```

Pour $n = 20$, on a une médiane moyenne de 1.93 et une EQM moyenne de 3,24%. L'EQM est élevée, ce qui indique une variabilité assez forte de la médiane par rapport à la vraie valeur de θ (2), et est dû à la faible taille de l'échantillon. La médiane moyenne est proche de la vraie valeur de θ, mais elle est également assez variable.

Pour $n = 100$, on a une médiane moyenne de 2.03 et une EQM moyenne de 0,61% L'EQM est considérablement réduite par rapport à $n = 20$, ce qui indique une réduction de la variabilité de la médiane par rapport à la vraie valeur de θ.

Pour $n = 1000$, on a une médiane moyenne de 2.01 et une EQM moyenne de 0,06% L'EQM est encore très réduite par rapport à $n = 20$ et $n = 100$, on se rapproche très fortement de la vraie valeur de θ !