

Tp Statistiques 2

Lapu Matthias | Amaël Kreis

Simulation et convergence

I. Variation sous-jacente et échantillonnage répété

1. Si $X \sim \mathcal{E}(0.5)$, quelle est la probabilité qu'on observe une valeur supérieure à 3?

On a :

$$f(x) = 0.5e^{-0.5x}$$

Donc pour :

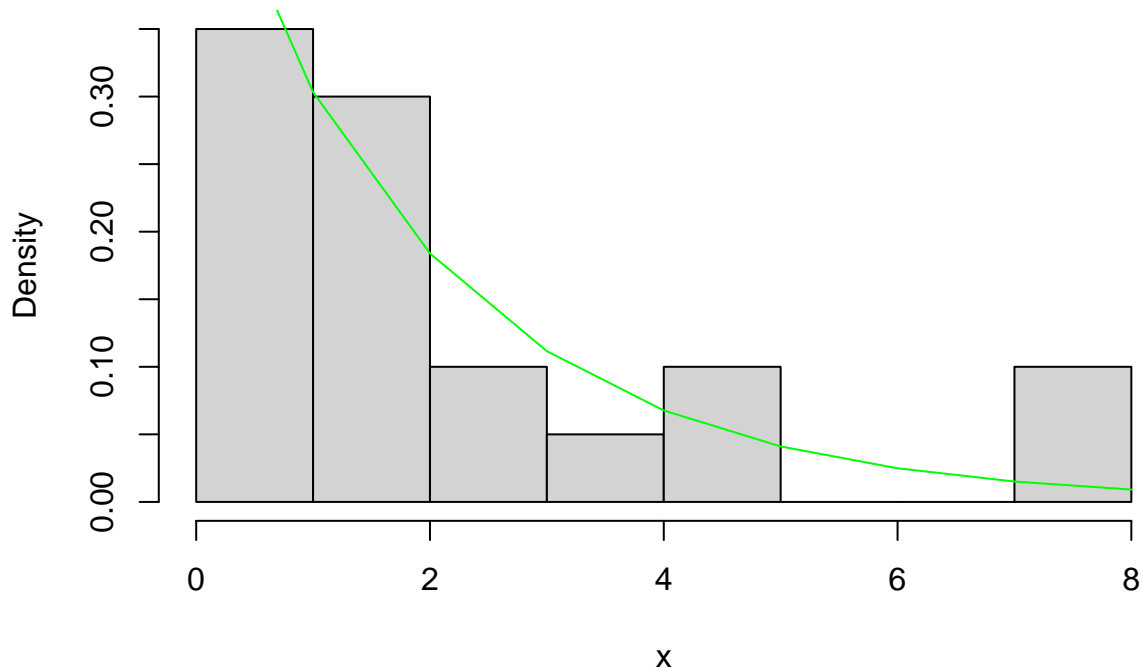
$$X \sim \mathcal{E}(0.5)$$

On trouve:

$$P(X > 3) = \int_3^{+\infty} \frac{1}{2} e^{-\frac{x}{2}} dx = [-e^{-\frac{x}{2}}]_3^{+\infty} = e^{-\frac{3}{2}} \approx 0.223$$

2. Simulez un échantillon de taille $n = 20$ d'une loi de $\mathcal{E}(0.5)$, créez un histogramme de votre échantillon et commentez la forme de votre histogramme. Superposez la vraie densité. Quelle est la probabilité empirique qu'on observe une valeur supérieure à 3 ?

Loi exponentielle pour $n = 20$



```
## [1] "Ici P(X>3)= 0.25"
```

L'histogramme peut être très proche de la densité ou au contraire s'en éloigner énormément.

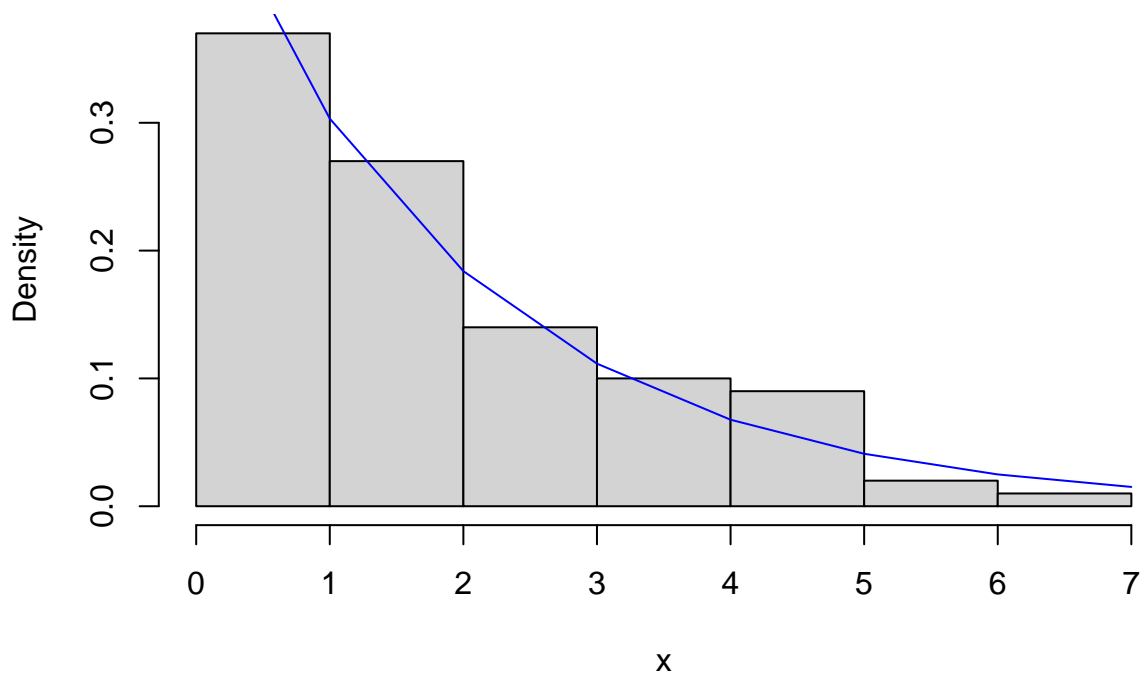
3. Répétez cette opération 5 ou 6 fois et commentez les différences entre les histogrammes que vous obtenez à chaque fois. Utilisez la même limite sur les axes pour faciliter la comparaison. Notez également comment la probabilité empirique qu'on observe une valeur supérieure à 3 change.

Les histogrammes sont tous très différents d'un échantillon à un autre, cela est dû à la faible taille des échantillons.

```
## [1] "Echantillon 1 P(X>3)= 0.3"
## [1] "Echantillon 2 P(X>3)= 0.15"
## [1] "Echantillon 3 P(X>3)= 0.4"
## [1] "Echantillon 4 P(X>3)= 0.15"
## [1] "Echantillon 5 P(X>3)= 0.5"
## [1] "Echantillon 6 P(X>3)= 0.1"
```

4. Augmentez la taille de votre échantillon à 100 et répétez votre expérience. Que remarquez-vous?

Loi exponentielle pour $n = 100$



est bien plus proche de la densité que précédemment.

II. Variabilité aléatoire du maximum de l'échantillon

1. Simuler un échantillon de taille $n = 10$ d'une loi $U(-1, 1)$ et enregistrez le maximum de l'échantillon.

```
loiU <- runif(10, -1, 1)
max1 <- max(loiu)
print(max1)
```

```
## [1] 0.7742051
```

2. Répétez les deux étapes ci-dessus dix fois, en écrivant le maximum de l'échantillon à chaque fois. Commentez la variabilité des valeurs que vous obtenez pour les maxima de votre échantillon.

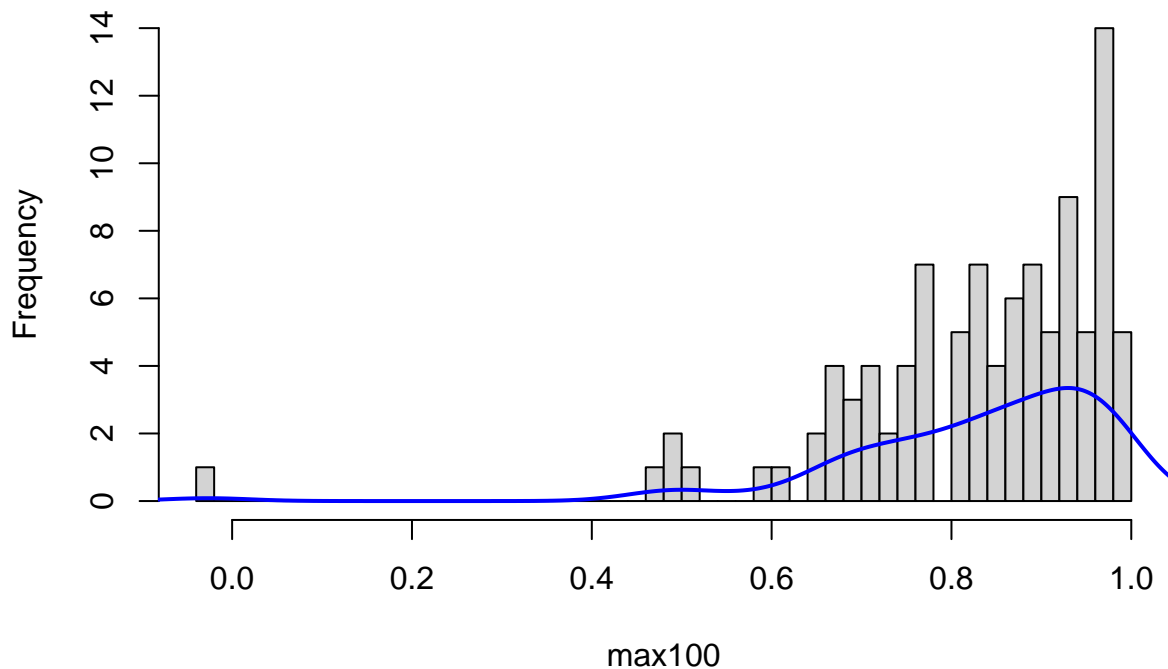
```
max10 <- c()
for (i in 1:10) {
  max10[i] = max(runif(10,-1,1))
}
print(max10)
```

```
## [1] 0.7371595 0.7976589 0.4453067 0.8679718 0.9473273 0.5736462 0.9502684
## [8] 0.3017782 0.9820409 0.9861358
```

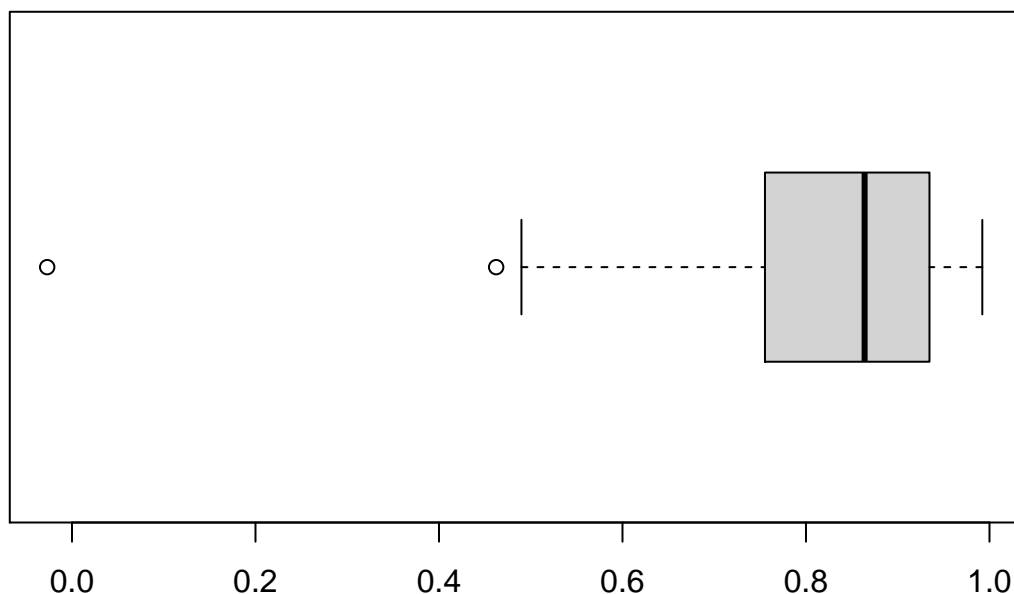
3. Répétez 100 fois et construisez un histogramme et une boîte à moustaches. Quelle est la loi du maximum, $M = \max 1 \leq i \leq n$ où $X_i \sim U(-1, 1)$ (TD1) ? Superposer la densité théorique sur l'histogramme. Que remarquez-vous ?

```
max100 <- c() #création d'un vecteur vide
for(i in 1:100){
  max100[i] = max(runif(10,-1,1)) #chaque maximum est inséré à la position i du vecteur
}
hist(max100,breaks=50)
densitermax100 <- density(max100) #cette fonction permet d'obtenir la densité de la loi
lines(densitermax100,lwd=2,col="blue") #superposition de la densité i
```

Histogram of max100



```
boxplot(max100, horizontal = TRUE)
```



On cherche à déterminer la loi du maximum d'un échantillon de loi uniforme :

$$F_M(x) = P(M \leq x) = P(\max(X_i) \leq x) = P(\{X_1 \leq x\} \cap \{X_2 \leq x\} \dots \{X_n \leq x\})$$

Par indépendance des X_i

$$= \prod_{i=1}^n P(X_i \leq x)$$

Car les X_i sont identiquement distribués, ils possèdent la même loi donc la même fonction de répartition.

$$= (F_{X_1}(x))^n$$

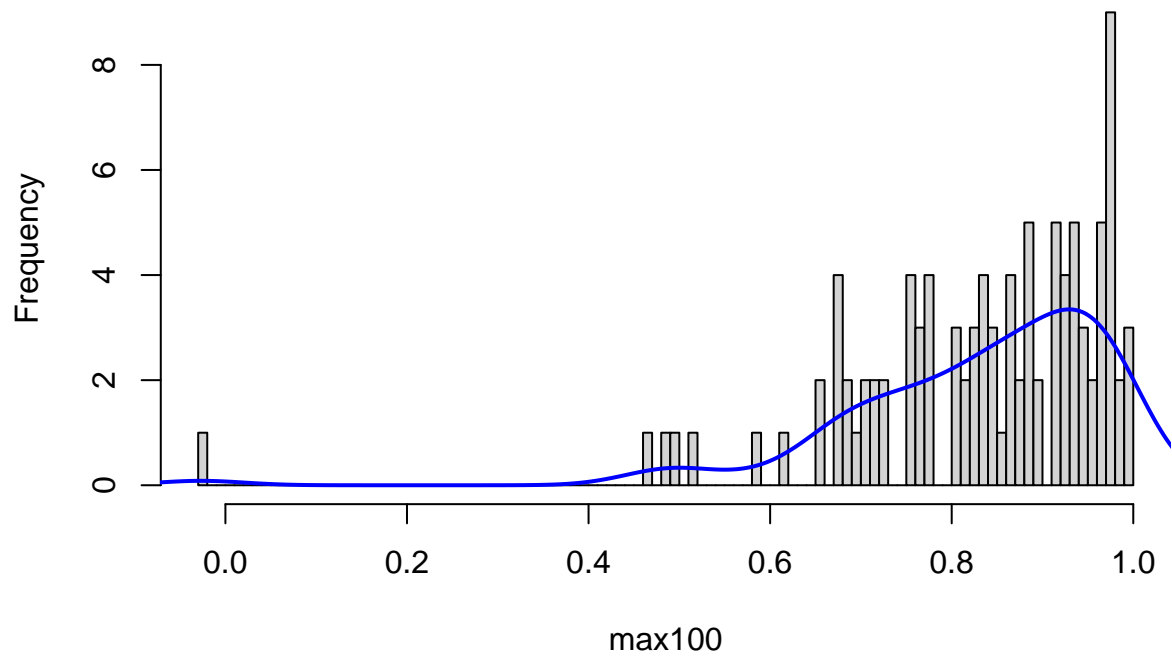
La loi du maximum, lorsque $X \in [a; b]$ est donc :

$$\left(\frac{x-a}{b-a}\right)^n$$

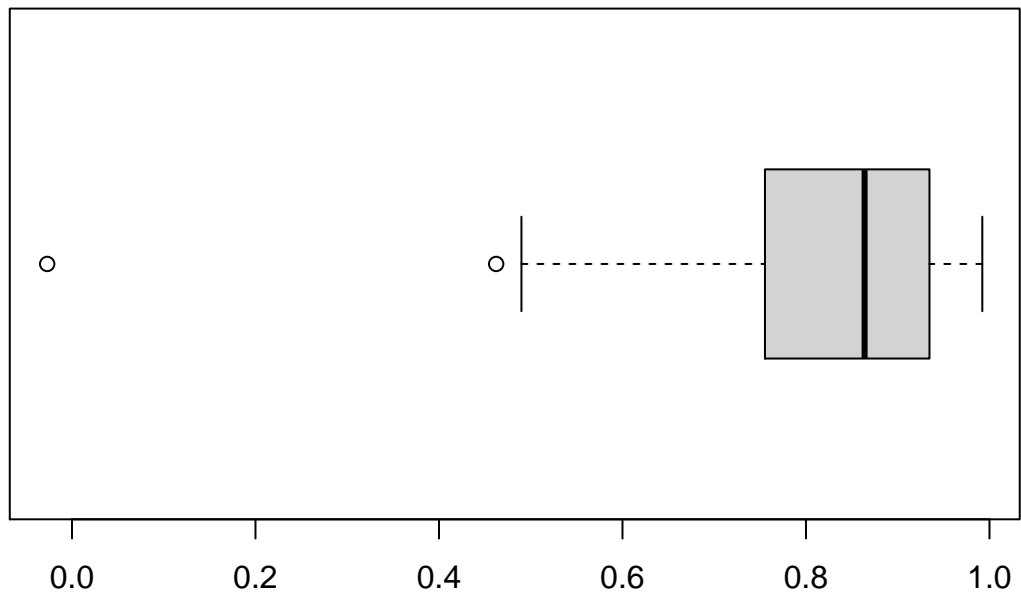
4. Augmentez la taille de votre échantillon à 50 et répétez votre expérience. Que remarquez-vous? Sont-ils proches de la symétrie ?

```
max100loi50 <- c()
for(i in 1:100){
  max100loi50[i] <- max(runif(50, -1, 1))
}
hist(max100, breaks=100)
densitermax100 <- density(max100)
lines(densitermax100, lwd=2, col="blue") #superposition de la densité de la loi
```

Histogram of max100



```
boxplot(max100, horizontal = TRUE)
```



Monte Carlo Methods

Vérifier que :

$$E[\hat{\theta}] = \theta$$

Nous utiliserons la linéarité de l'espérance :

$$E\left[\frac{1}{n} \sum_{i=1}^n \psi(X)\right] = \frac{1}{n} \sum_{i=1}^n E[\psi(X)] = \frac{1}{n} * n * \theta = \theta$$

Moyenne et phénomène de concentration.

1. Donner une borne de cette quantité en utilisant l'inégalité de Bienaymé Chebychev.

Rappelons l'inégalité de Bienaymé Chebychev :

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

Calculons la variance grâce à son caractère quadratique ainsi qu'à l'indépendance des X_i :

$$V(\hat{\theta}) = V\left(\frac{1}{n} \sum_{i=1}^n \psi(X_i)\right) = \frac{1}{n^2} V\left(\sum_{i=1}^n \psi(X_i)\right) = \frac{1}{n^2} \sum_{i=1}^n V(\psi(X_i)) = \frac{1}{n^2} * n\sigma^2 = \frac{\sigma^2}{n}$$

On retrouve donc une borne pour cette inégalité.

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

2. En supposant que $a \leq (X_i) \leq b$, donner une borne en utilisant l'inégalité de Hoeffding.

Posons :

$$S_n = \sum_{k=1}^n \psi(X_k)$$

D'après l'énoncé, nous savons que :

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

Ainsi :

$$\frac{S_n}{n} = \hat{\theta}$$

Donc :

$$S_n = n\hat{\theta}$$

D'après l'inégalité de Hoeffding nous savons que :

$$P(|S_n - E(S_n)| \geq t) \leq 2\exp\left(\frac{-2t^2}{\sum_{k=1}^n (b_k - a_k)^2}\right)$$

Calculons l'espérance de S_n :

$$E(S_n) = E\left(\sum_{k=1}^n \psi(X_k)\right) = \sum_{k=1}^n E(\psi(X_k)) = \sum_{k=1}^n \theta = n\theta$$

Ainsi :

$$P(|n\hat{\theta} - n\theta| \geq t) \leq 2\exp\left(\frac{-2t^2}{\sum_{k=1}^n (b_k - a_k)^2}\right) P(|\hat{\theta} - \theta| \geq \frac{t}{n}) \leq 2\exp\left(\frac{-2t^2}{\sum_{k=1}^n (b_k - a_k)^2}\right)$$

On pose :

$$\frac{t}{n} = \delta P(|\hat{\theta} - \theta| \geq \delta) \leq 2\exp\left(\frac{-(2n\delta)^2}{\sum_{k=1}^n (b_k - a_k)^2}\right)$$

3. De combien d'échantillons auriez-vous besoin pour que la probabilité que $\hat{\theta} = \theta$ soit inférieur à 1%.

Servons-nous de l'inégalité de Bienaymé-Tchebychev que nous avons trouvé lors du 1.

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

Remplaçons par :

$$\delta = 2\sigma\delta^2 = 4\sigma^2$$

Ainsi :

$$\frac{\sigma^2}{n\delta^2} = \frac{1}{4n}$$

Or l'énoncé demande à ce que la probabilité soit inférieure à 1% , c'est-à-dire que :

$$\frac{1}{4n} = 0.01 \implies n = 25$$

Afin que la probabilité soit inférieure à 1%, il faut 25 échantillons.

Application pour l'estimation de probabilité

1. Pour la question 1 :

L'espérance de la loi est 2. En prenant un échantillon

$$(X_1, X_2, \dots, X_n)$$

Le paramètre d'intérêt est donc l'estimateur de cette loi:

$$X_i \sim \varepsilon(\theta) \forall \theta \in R_+^* E[X_i] = \frac{1}{\theta} \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

La moyenne empirique tend presque surement vers l'estimateur. Ainsi :

$$\hat{\theta}_n = \frac{1}{\bar{X}}$$

2. Utilisons l'inégalité de Hoeffding.

L'énoncé demande à ce que $E(Z) =$ garantie probabiliste de l'erreur. Il faut trouver Z .

On pose :

$$Z = 1_{|\hat{\theta}_n - \theta| \geq \delta}$$

Ainsi, nous avons donc :

$$\eta = P(|\hat{\theta}_n - \theta| \geq \delta) = E[Z]$$

Le paramètre d'intérêt est la moyenne de la fonction exponentielle, c'est à dire 2.

La moyenne empirique tend presque surement vers la moyenne , ainsi l'estimateur est donc consistant. Il faut ensuite utiliser l'inégalité de Hoeffding.

$$\eta \approx \frac{1}{n} \sum_{i=1}^n Z_i$$

On va donc essayer d'estimer cela à l'aide de R :

Théorème Central Limite et Estimation Monte Carlo

1. Vérifier que l'espérance théorique d'une loi de Pareto est $E[X] = a/(\alpha - 1)$.

$$P(X \leq t) = (1 - (\frac{a}{t})^\alpha), \text{ avec } t \geq a$$

Donc :

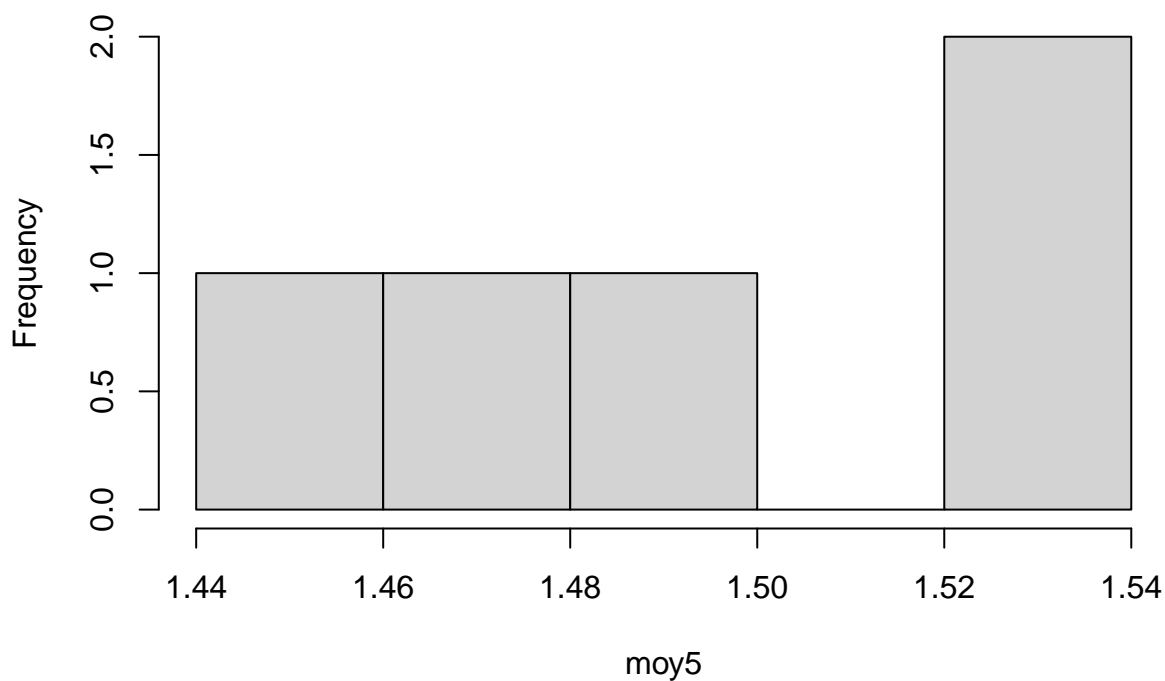
$$E(X) = \int_0^{+\infty} 1 - P(X \leq t) dt = \int_0^{+\infty} P(X > t) dt = a + a^\alpha \int_a^{+\infty} \frac{1}{t^\alpha} dt = a + \frac{a}{\alpha - 1} = \frac{\alpha a}{\alpha - 1}$$

2 et 3 . Simuler $N = 1000$ échantillons i.i.d de loi commune Pareto $P(a, \alpha)$ (avec votre choix de paramètres) de taille $n = 5, 30, 100$ et calculer les moyennes et variances empiriques $\bar{X}_{n,i}$ et $S_{n,i}$, $i = 1, \dots, N$. Puis tracer l'histogramme des moyennes empiriques.

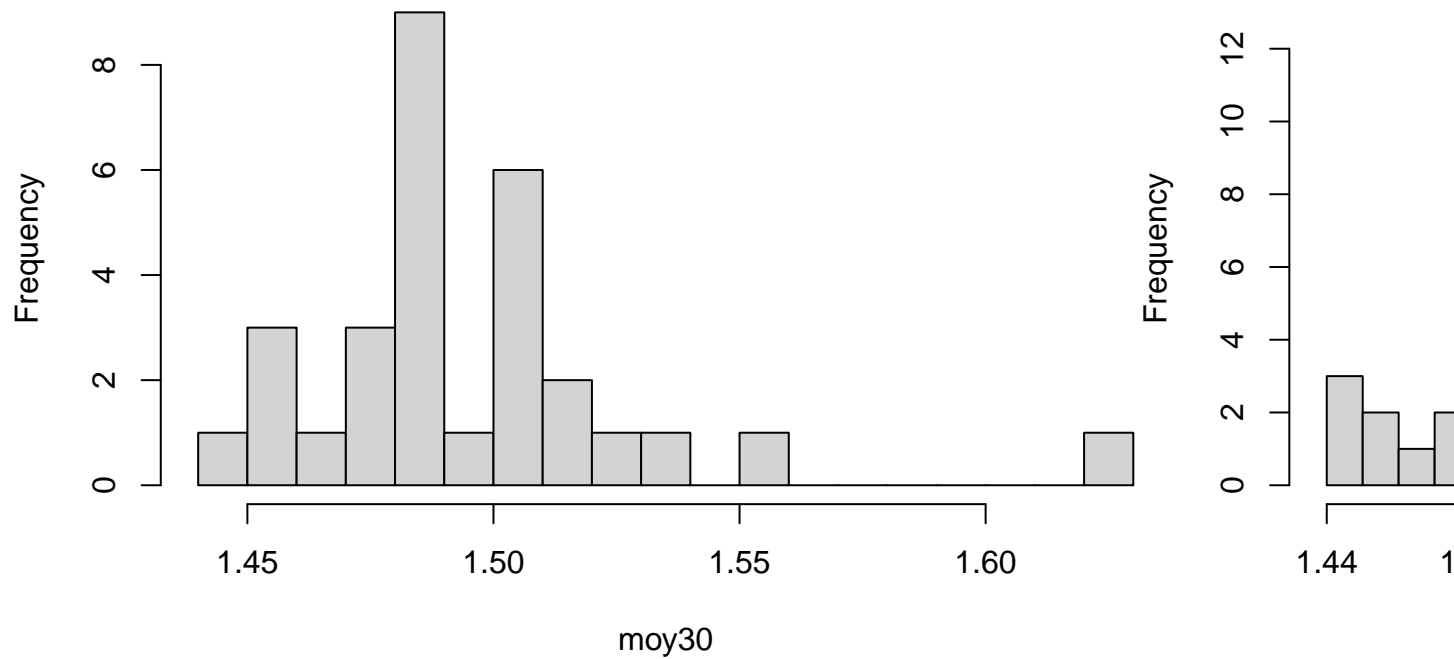
On prend

$$a = 1, \alpha = 3$$

Moyennes empiriques pour n=5

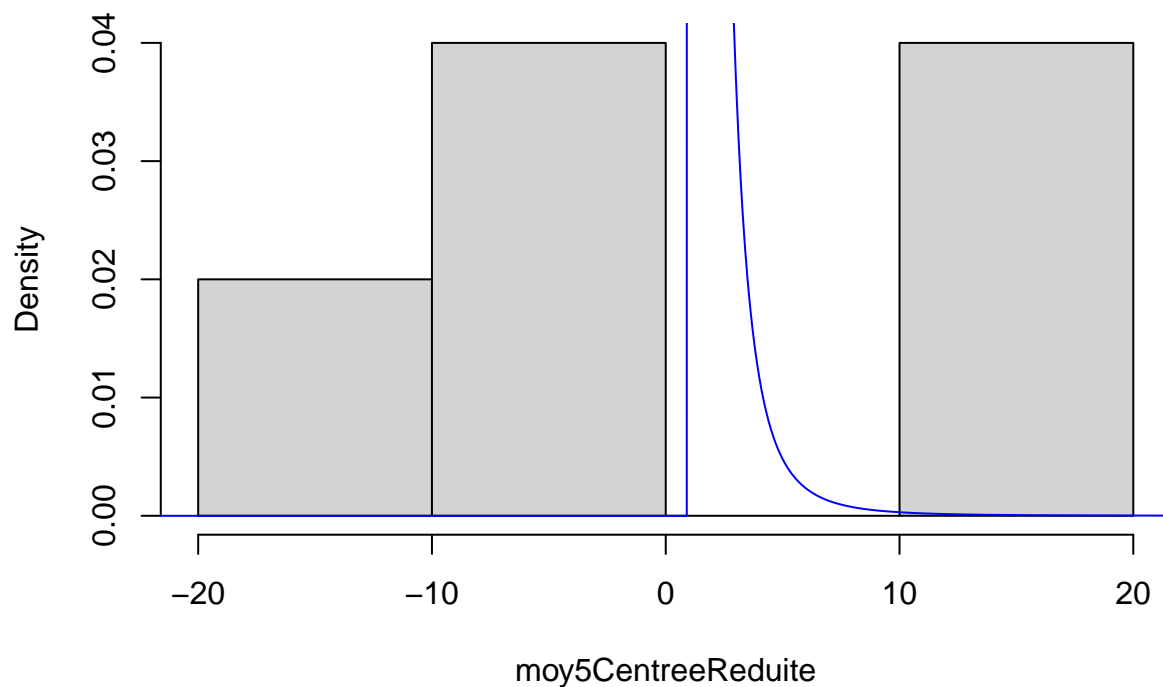


Moyennes empiriques pour n=30

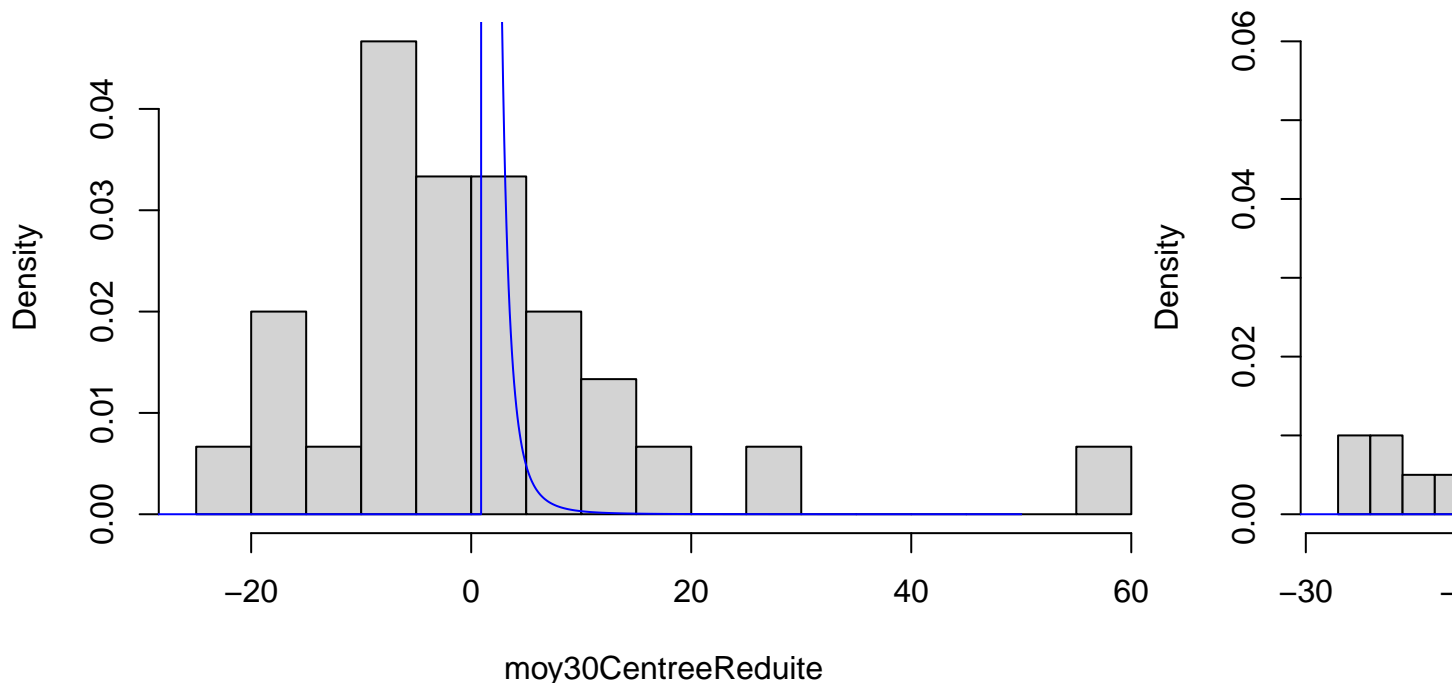


4. A l'aide d'une renormalisation adéquate (a_n , b_n), montrer que $U_{n,i} = \frac{\bar{X}_n - a_n}{b_n}$ a une loi que vous pouvez approcher. Comparez histogramme de les moyennes empiriques normalisées, $U_{n,i}$, et distribution théorique approchée. Quelle est l'influence de la taille de l'échantillon n sur la qualité de cette approximation?

Histogram of moy5CentreeReduite



Histogram of moy30CentreeReduite



Quand le théorème de central limite ne s'applique pas

1. Simuler un échantillon de taille $n = 20$ d'une loi de $C(2)$ et calculer la moyenne empirique \bar{X}_n .

Cauchy20 est une simulation d'un échantillon $n=20$.

```
## [1] 9.08703
```

2. Faites varier la taille de l'échantillon $n = 20, 100, 1000$ et 10000 . Qu'en déduire ?

```
## [1] 9.08703
```

```
## [1] -0.1928555
```

```
## [1] 0.5853447
```

```
## [1] 1.659902
```

On remarque que malgré le nombre élevé de l'échantillon la moyenne ne semble pas se stabiliser comme pour une loi normale.

3. Expliquer ce comportement

Nous savons d'après le cours de probabilités que la loi de cauchy n'admet pas d'espérance ni d'écart type. Cela explique donc le comportement de la moyenne malgré la taille de l'échantillon.

4. Quelle est la médiane d'une loi de cauchy ?

La courbe est symétrique, la médiane d'une loi de cauchy est θ . D'après le manuel de R, quand la position n'est pas défini celle-ci est mise à 0. Par conséquent nous devons vérifier si la médiane semble proche de 0.

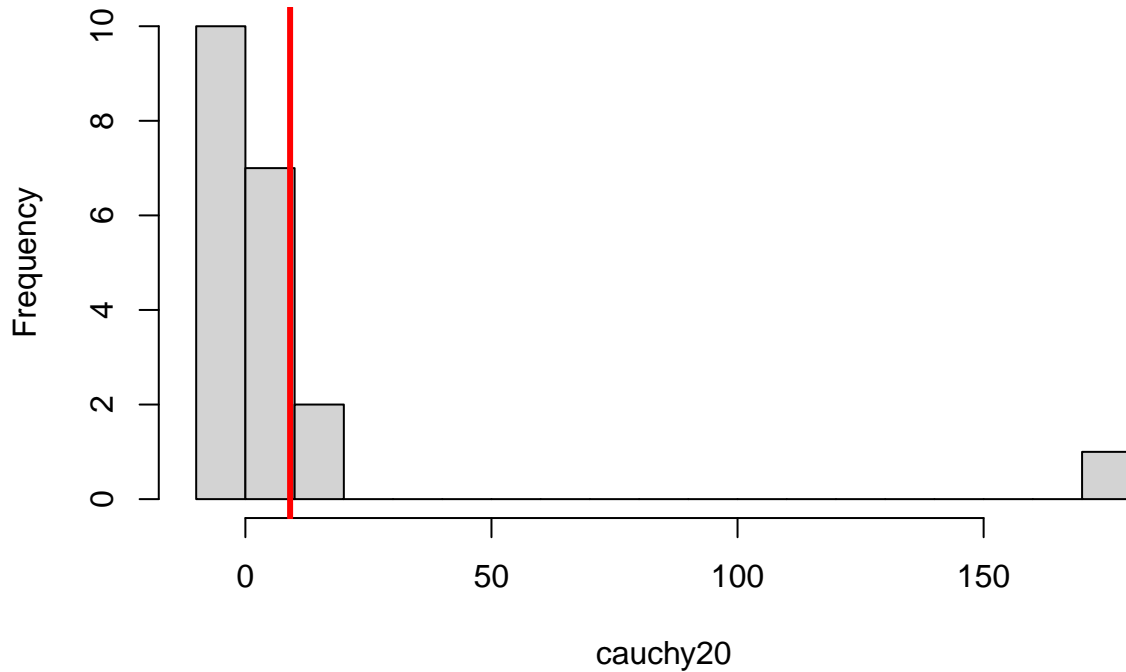
$$f(x, \theta) = \frac{1}{\pi} \frac{1}{1 + (x - \theta)^2} \frac{1}{2} = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{dx}{1 + (x - \theta)^2} F^{-1}\left(\frac{1}{2}\right) = \theta$$

5. En déduire un estimateur de θ et évaluer la performance de cet estimateur sur les différents échantillons.

Nous pouvons essayer d'approximer θ , c'est-à-dire la médiane, cela revient donc à chercher une estimation du quantile en 0.5. D'après le cours, les quantiles permettent de localiser les valeurs les plus fréquentes. Nous allons donc essayer d'estimer le quantile.

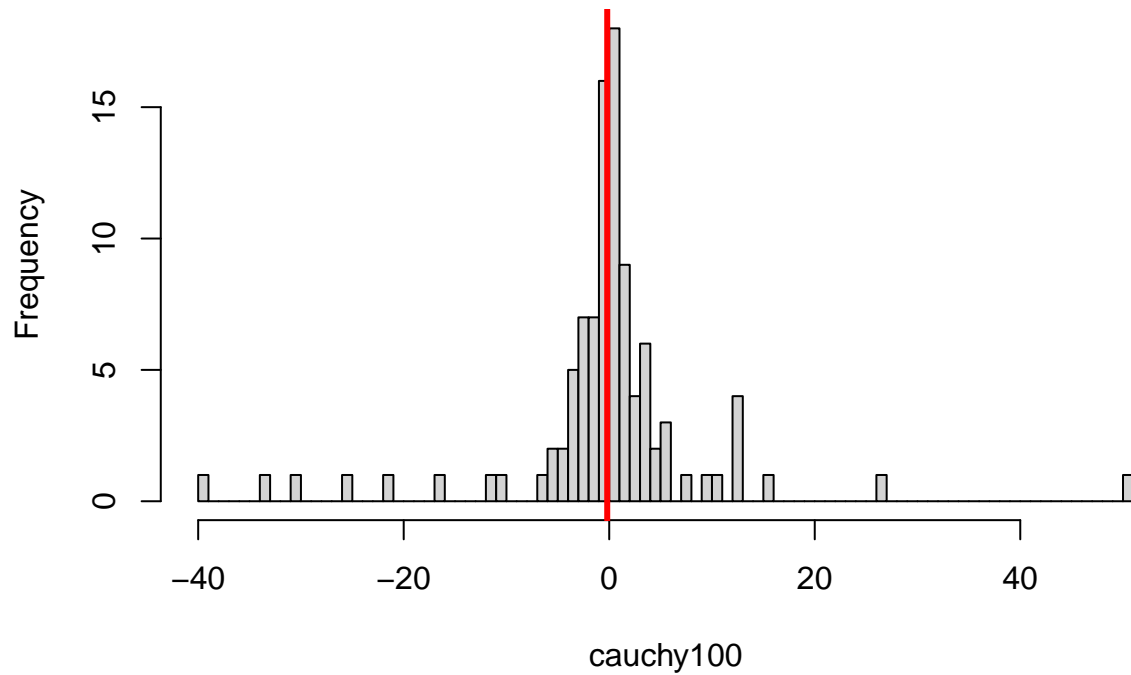
```
hist(cauchy20,breaks=20)
abline(v=mean(cauchy20),col="red",lwd=3)
```

Histogram of cauchy20



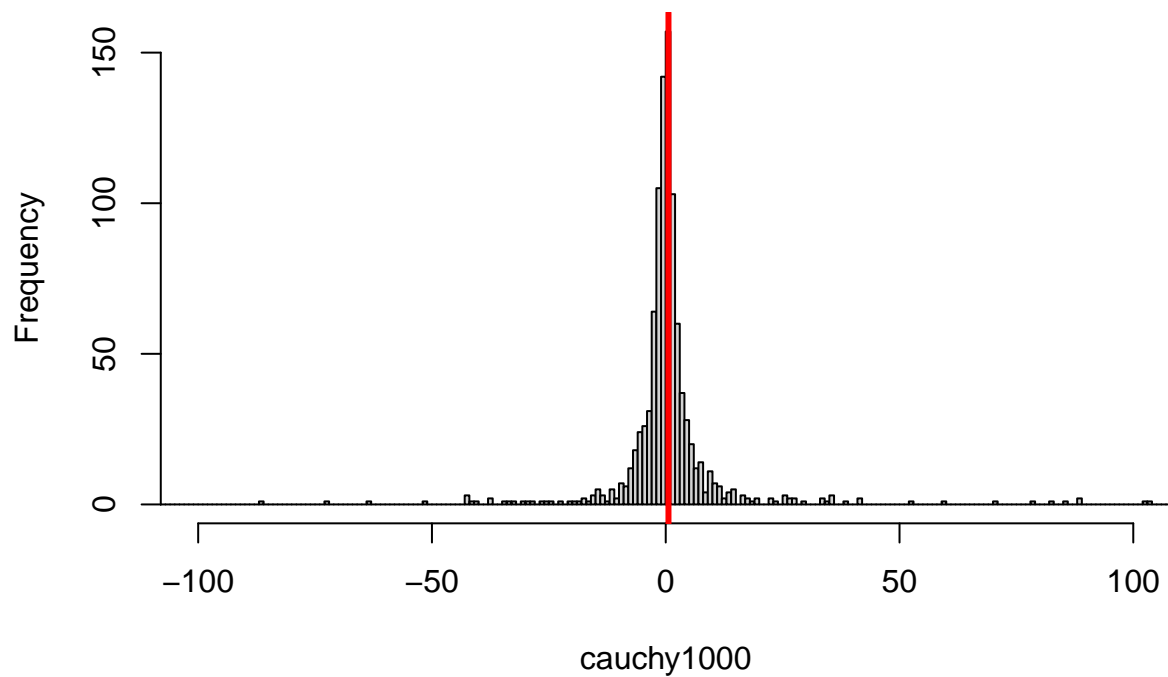
```
hist(cauchy100,breaks=100)
abline(v=mean(cauchy100),col="red",lwd=3)
```

Histogram of cauchy100



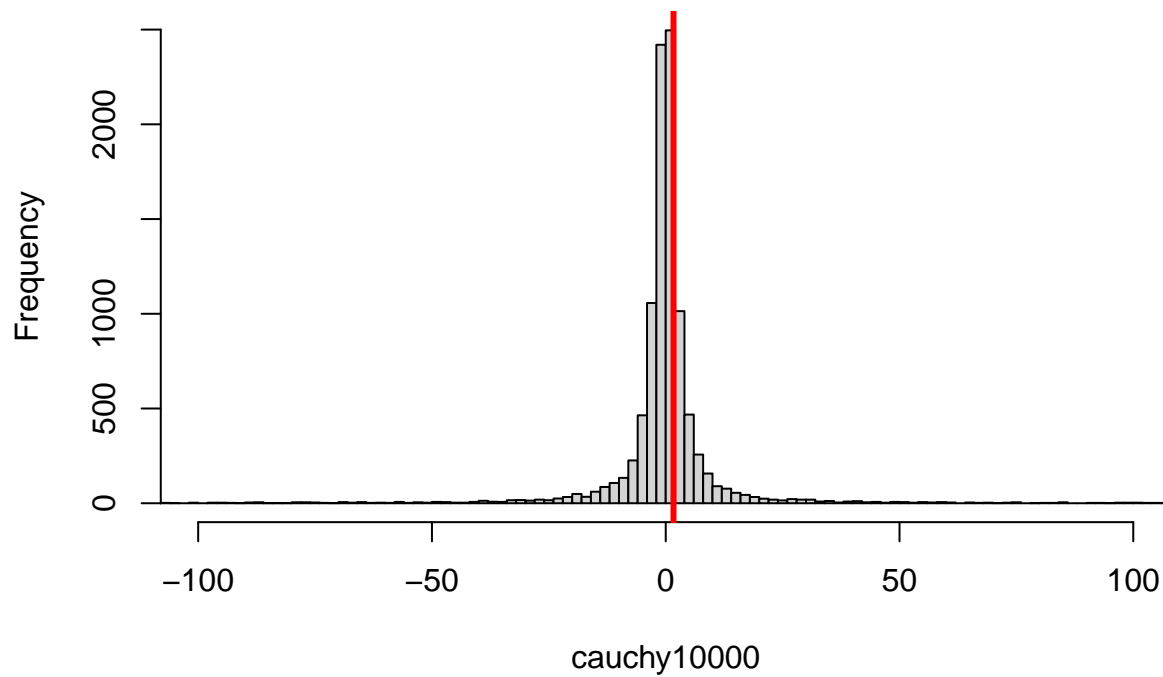
```
hist(cauchy1000,xlim=c(-100,100),breaks=1000)  
abline(v=mean(cauchy1000),col="red",lwd=3)
```

Histogram of cauchy1000



```
hist(cauchy10000,xlim=c(-100,100),breaks=10000)  
abline(v=mean(cauchy10000),col="red",lwd=3)
```

Histogram of cauchy10000



D'après les graphiques nous pouvons remarquer que l'estimation de theta semble proche de la vrai valeur, n'ayant pas de moyen de calculer l'espérance de la loi cela semble être un bon estimateur, car celui-ci semble proche de 0, assez pour jugé les performances de cet estimateur comme suffisant.