Tp Statistiques 2

Lapu Matthias | Amaël Kreis

Simulation et convergence

I. Variation sous-jacente et échantillonnage répété

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

On a :

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

Donc pour :

$$X \sim \epsilon(0.5)$$

On trouve:

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

2.Simulez un échantillon de taille n = 20 d'un loi de $\epsilon(0, 5)$, créez un histogramme de votre échantillon et commentez la forme de votre histogramme. Superposer la vrai densité. Quelle est la probabilité empirique qu'on observe une valeur supérieure à 3 ?

```
x<-rexp(20,0.5)
hist(x, freq=FALSE,main = "Loi exponentielle pour n = 20")
maxvalue <- ceiling(max(x))
lines(0:maxvalue,dexp(0:maxvalue, 0.5), col="red",)
```


```
print(paste("Ici P(X>3)=",sum(x>3)/20))
```

```
## [1] "Ici P(X>3)= 0.2"
```

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.

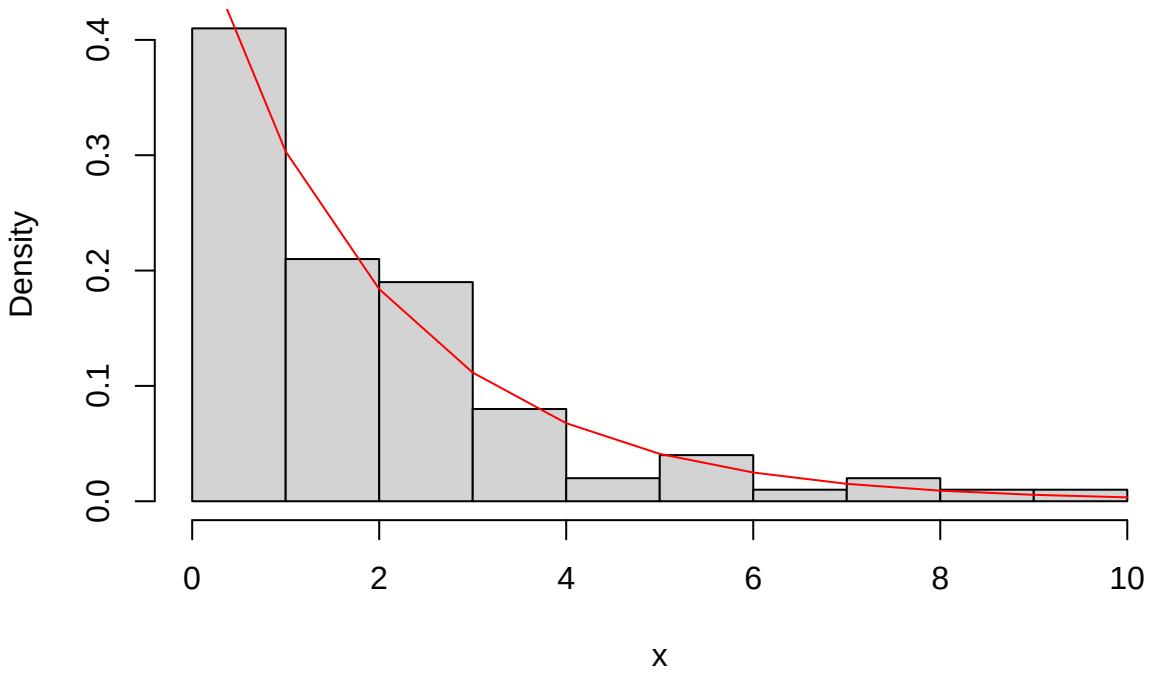
```
for (i in 1:6) {
  x<-rexp(20,0.5)
  print(paste("Echantillon ",i," P(X>3)=",sum(x>3)/20))
}
```

```
## [1] "Echantillon  1  P(X>3)= 0.3"
## [1] "Echantillon  2  P(X>3)= 0.05"
## [1] "Echantillon  3  P(X>3)= 0.4"
## [1] "Echantillon  4  P(X>3)= 0.25"
## [1] "Echantillon  5  P(X>3)= 0.05"
## [1] "Echantillon  6  P(X>3)= 0.05"
```

Les résultats sont tous très différent d'un échantillon à un autre, cela est du à la faible taille des échantillons.

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

```
x<-rexp(100,0.5)
hist(x, freq=FALSE,main = "Loi exponentielle pour n = 100")
maxvalue <- ceiling(max(x));
lines(0:maxvalue,dexp(0:maxvalue, 0.5), col="red",)
```



On remarque que l'histogramme est bien plus proche de la densité que précédement.

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] 0.2917475
```

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() #création d'un vecteur vide
for (i in 1:10) {
  max10[i] = max(runif(10,-1,1))#chaque maximum est inséré à la position i du vecteur
}
print(max10)
```

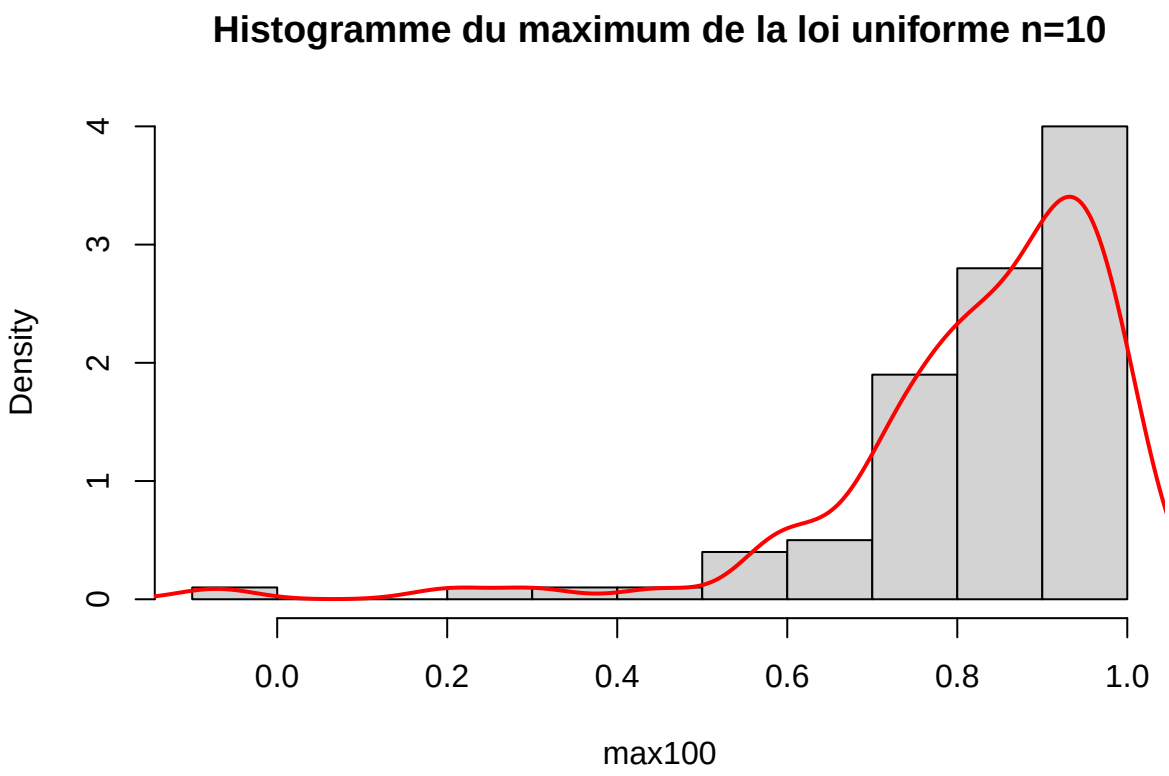
```
##  [1] 0.6619602 0.8522022 0.9724027 0.9194693 0.9724979 0.8527813 0.8936298
##  [8] 0.7762990 0.9823970 0.9305100
```

Les valeurs sont très variables, on remarque notamment qu'il est très rare d'avoir des valeur en dessous de 0.5 et que de nombreuse valeurs sont au dessus de 0.8.

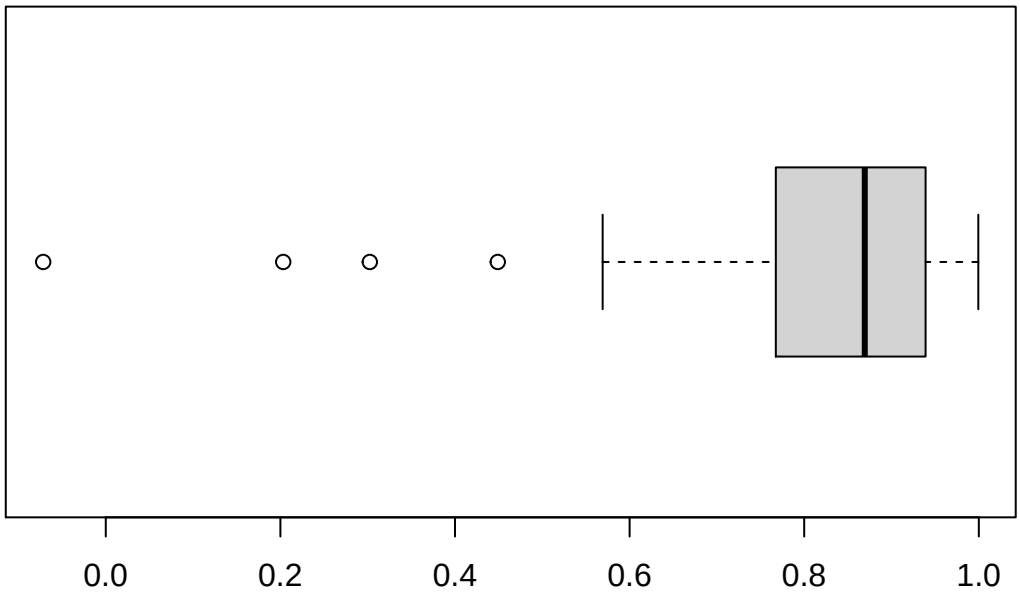
3.Répétez 100 fois et construisez un histogramme et une boîte à moustaches. Quelle est la loi du maximum, $M = max(1 \leq i \leq nX_i)$ où $X_i \sim U(-1, 1)$ (TD1) ? Superposer la densité théorique sur l'histogramme. Que remarquez-vous ?

```
max100 <- c()
for(i in 1:100){
```

```
 max100[i] = max(runif(10,-1,1))
}
hist(max100,breaks=10,main = "Histogramme du maximum de la loi uniforme n=10", freq = FALSE)
densitermax100 <- density(max100) #cette fonction permet d'obtenir la densité de la loi
lines(densitermax100,lwd=2,col="red") #superposition de la densité i
```




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



On remarque que la densité est très proche des maximums obtenus.

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

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

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.

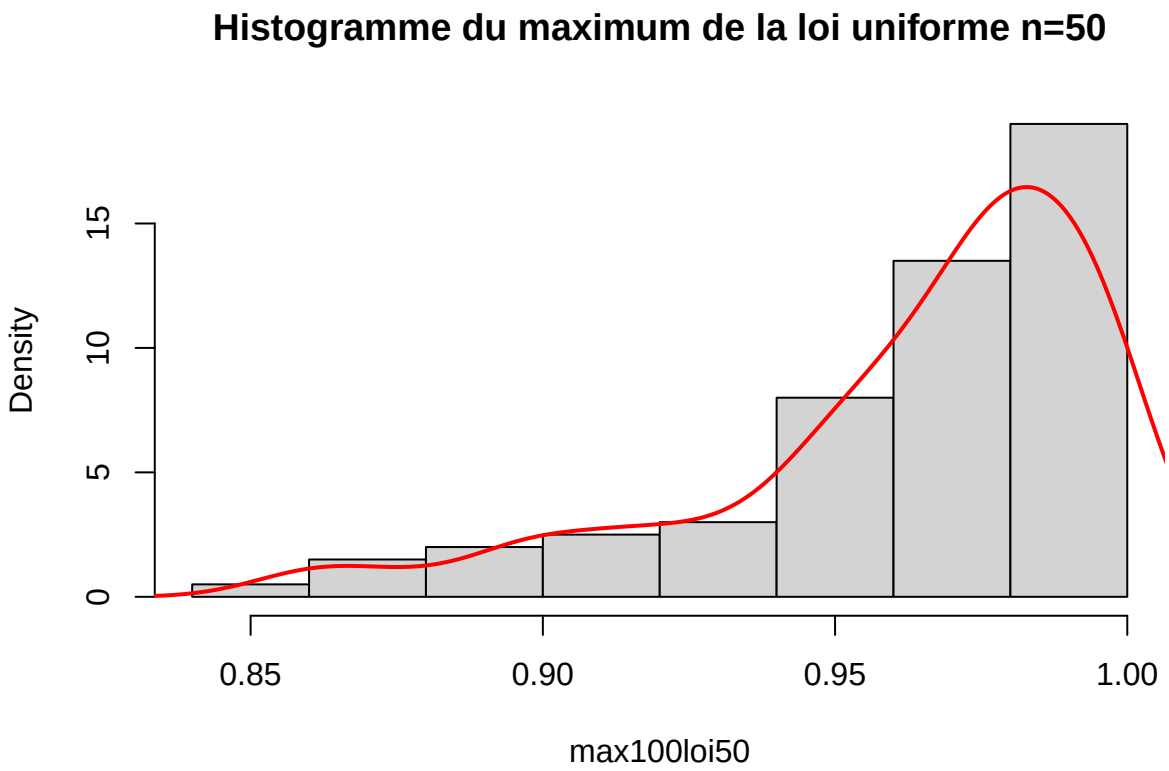
$$= (F_{X_1}(x))^n$$

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

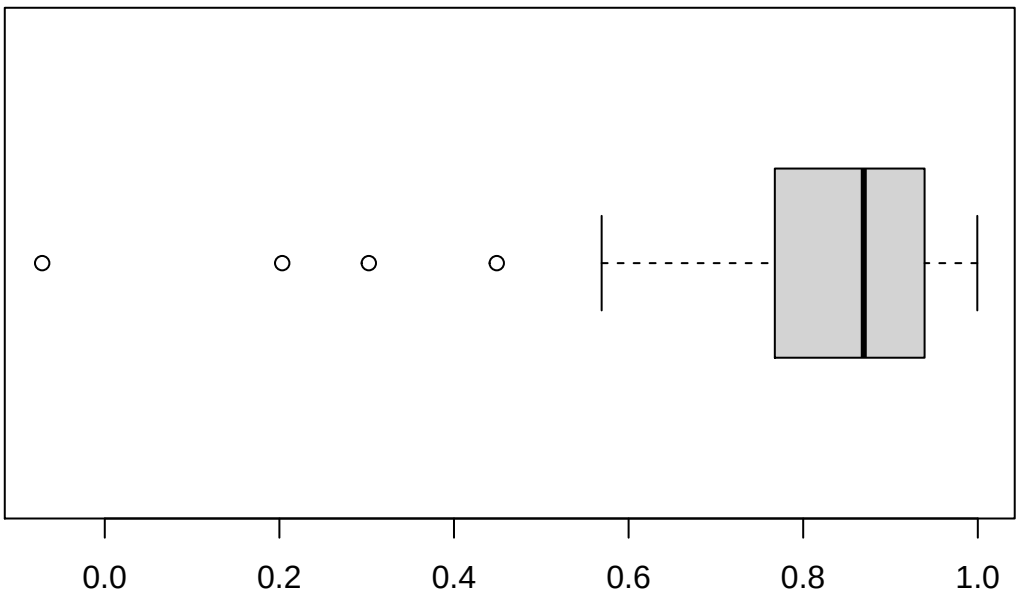
$$(\frac{x-a}{b-a})^n$$

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(max100loi50,breaks=10, main = "Histogramme du maximum de la loi uniforme n=50", freq = FALSE)
lines(density(max100loi50),lwd=2,col="red")
```




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



Ils ne semblent pas particulièrement plus proches de la symétrie.

Monte Carlo Methods

Vérifier que :

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

Nous utiliserons la linéarité de l'espérance :

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

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 de cette quantité grâce au caractère quadratique de la variance ainsi qu'à l'indépendance des X_i :

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

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 $a \leq \psi(Xi) \leq b$, 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 :

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

Calculons l'esperance de S_n:

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

Ainsi :

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

On pose :

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

3. De combien d'échantillons auriez-vous besoin pour que la probabilité que $\delta = 2\sigma$ 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 inférieur à 1% , c'est-à-dire que :

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

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

Application pour l'estimation de probabilité

1. Pour la question I avec $\epsilon(0.5)$, identifier le paramètre d'intérêt et donner un estimateur $\hat{}$.

On cherche un estimateur pour la loi exponentielle. En prenant un échantillon :

$$(X_1, X_2, ..., X_n)$$

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

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 :

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

Ainsi, nous avons donc :

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

D'après la méthode de Monte Carlo, on a donc :

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

Théorème Central Limite et Estimation Monte Carlo

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

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

Donc :

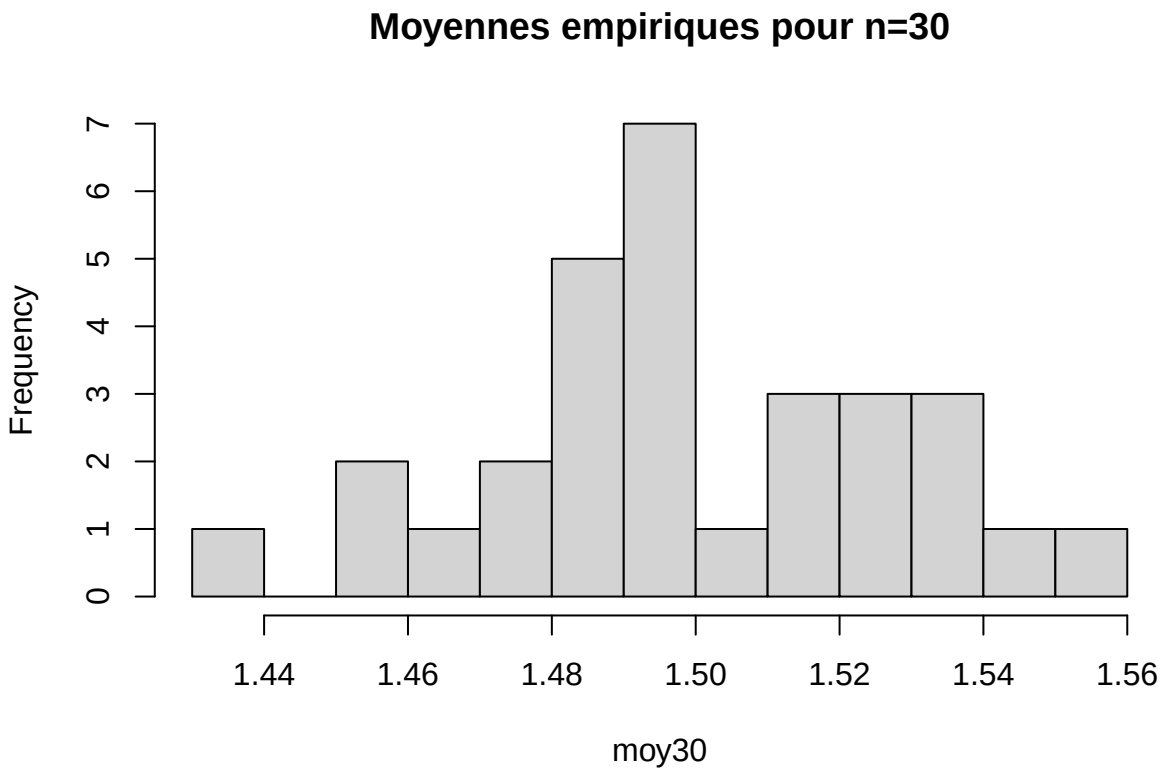
$$E(X) = \int_0^{+\infty} 1 - P(X \leq t)dt = \int_0^{+\infty} P(X > t)dt = a + a^{\alpha} \int_a^{+\infty} \frac{1}{t^{\alpha}} dt = a + \frac{a}{\alpha - 1} = \frac{\alpha a}{\alpha - 1}$$

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

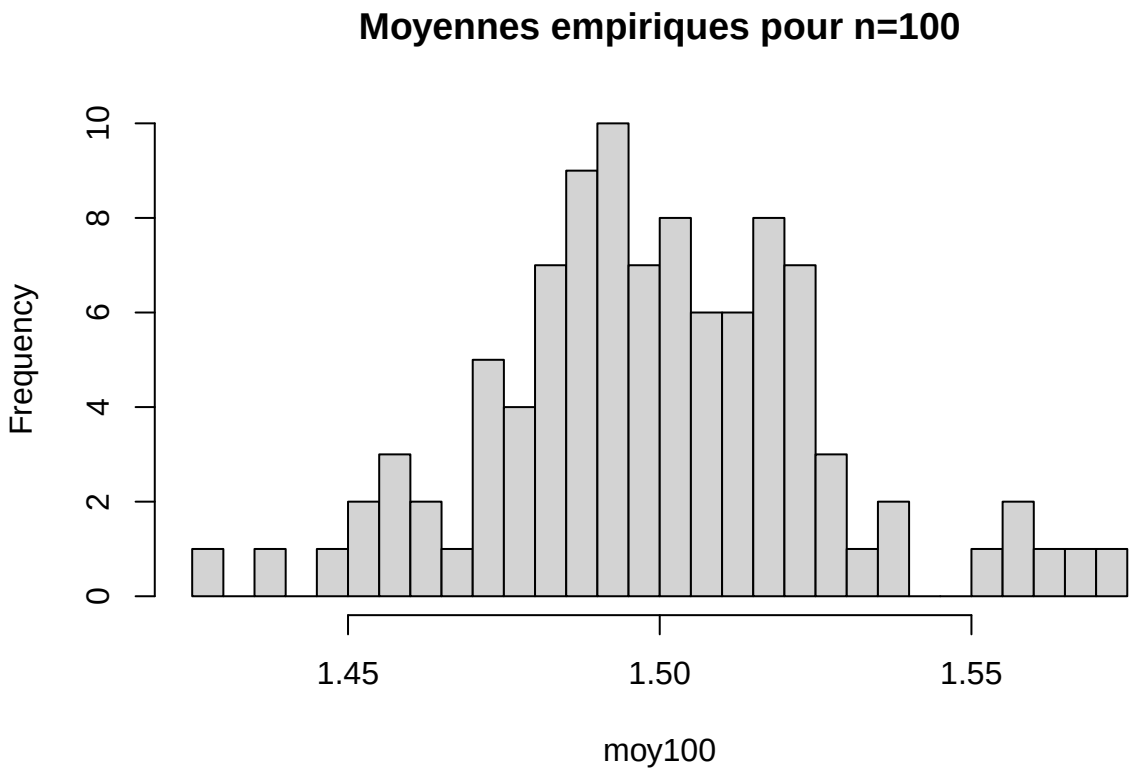
On prend

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

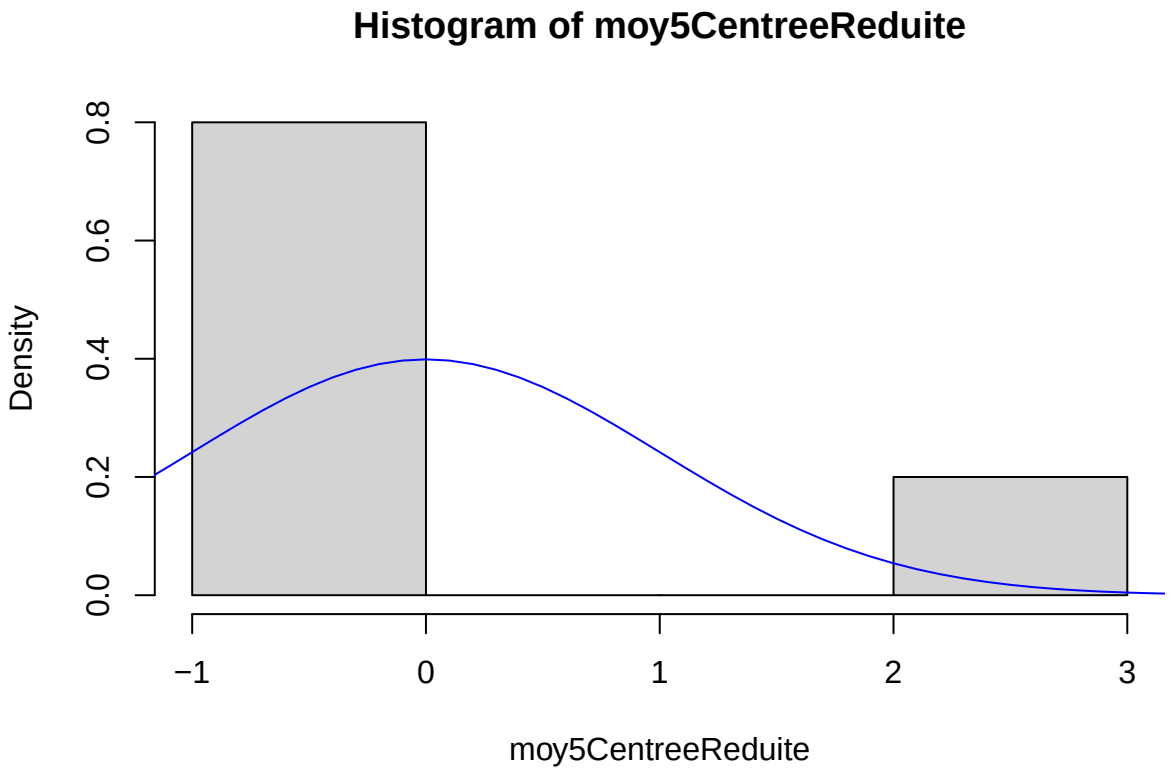

```r
hist(moy30,breaks = 15,main="Moyennes empiriques pour n=30")
```

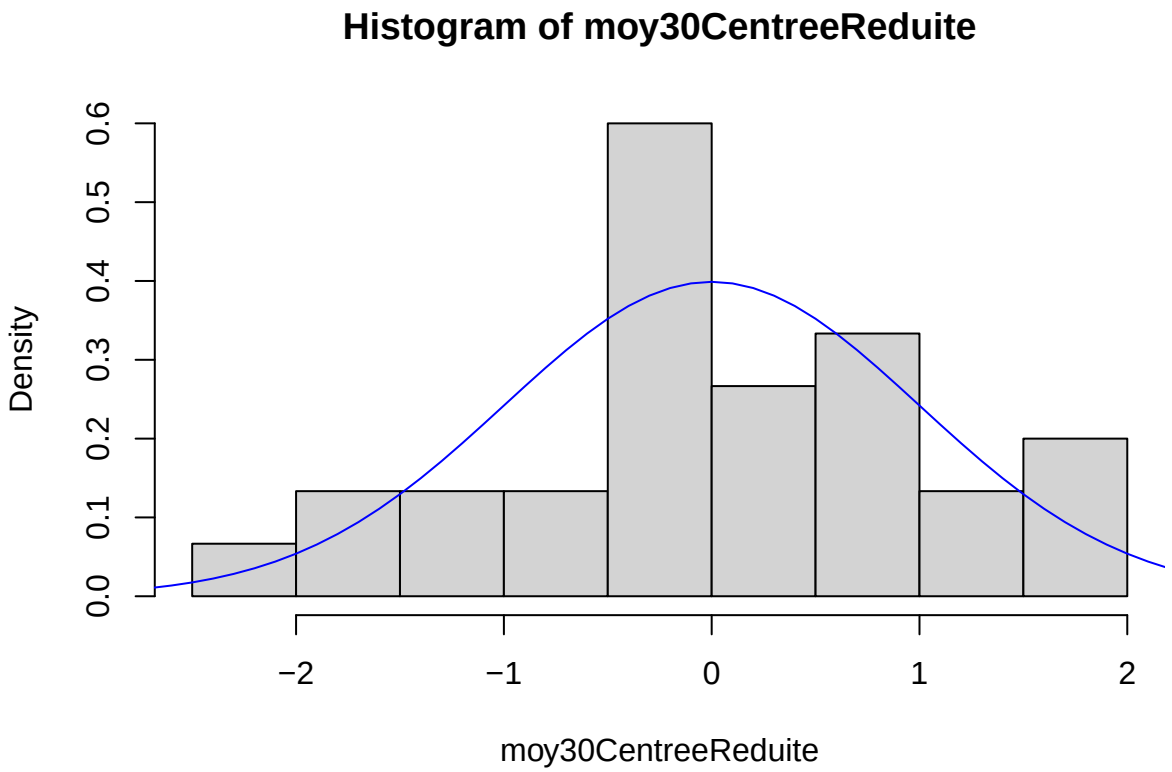


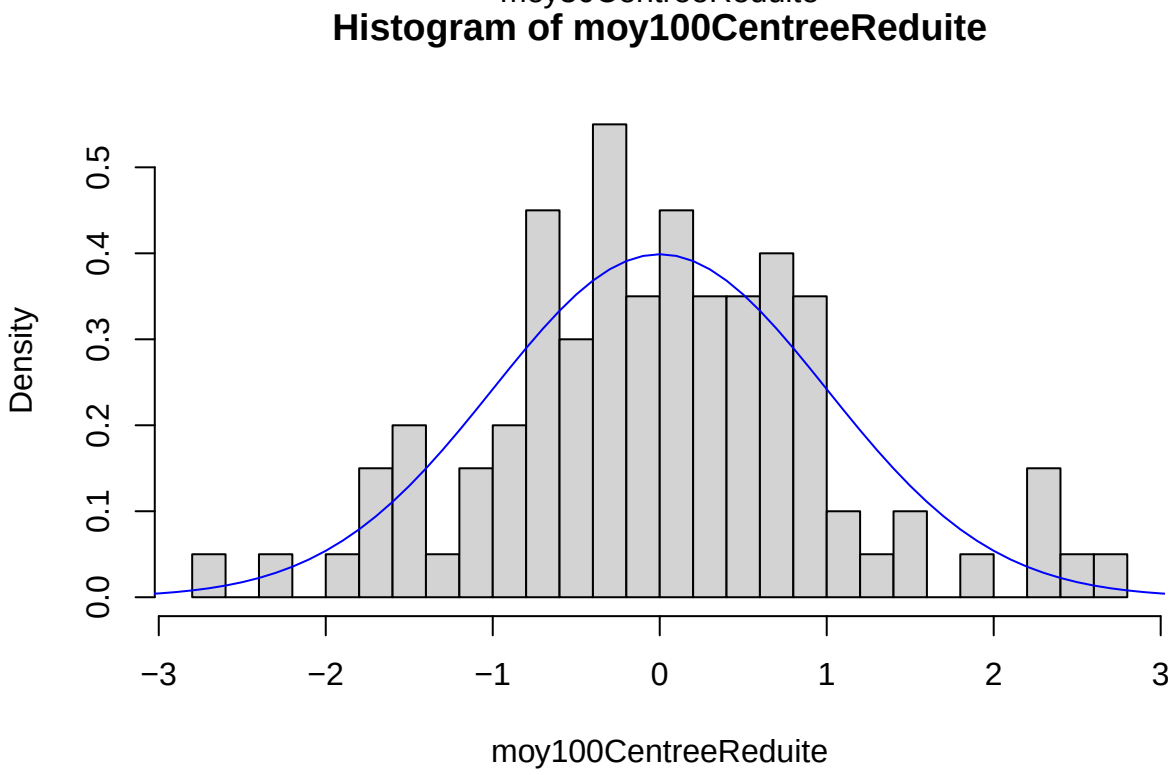

```r
hist(moy100,breaks = 30,main="Moyennes empiriques pour n=100")
```



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







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] 4.825236
```

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

```
## [1] 4.825236
```

```
## [1] 264.2845
```

```
## [1] 4.822681
```

```
## [1] -0.5054119
```

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 ?

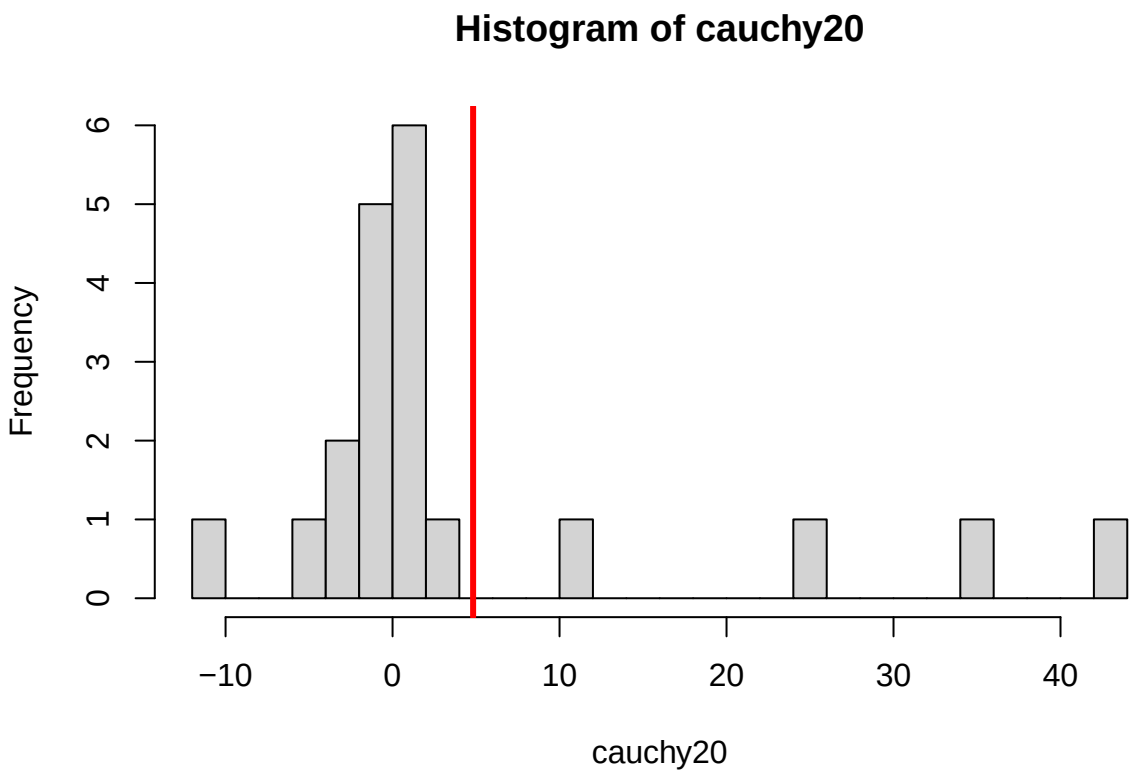
La courbe est symétrique , la médiane d'une loi de Cauchy est θ. D'après le manuel de R, quand la position n'est pas définie celle-ci est mise à 0. Par conséquent, nous devons vérifier si la médiane semble proche de 0.

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

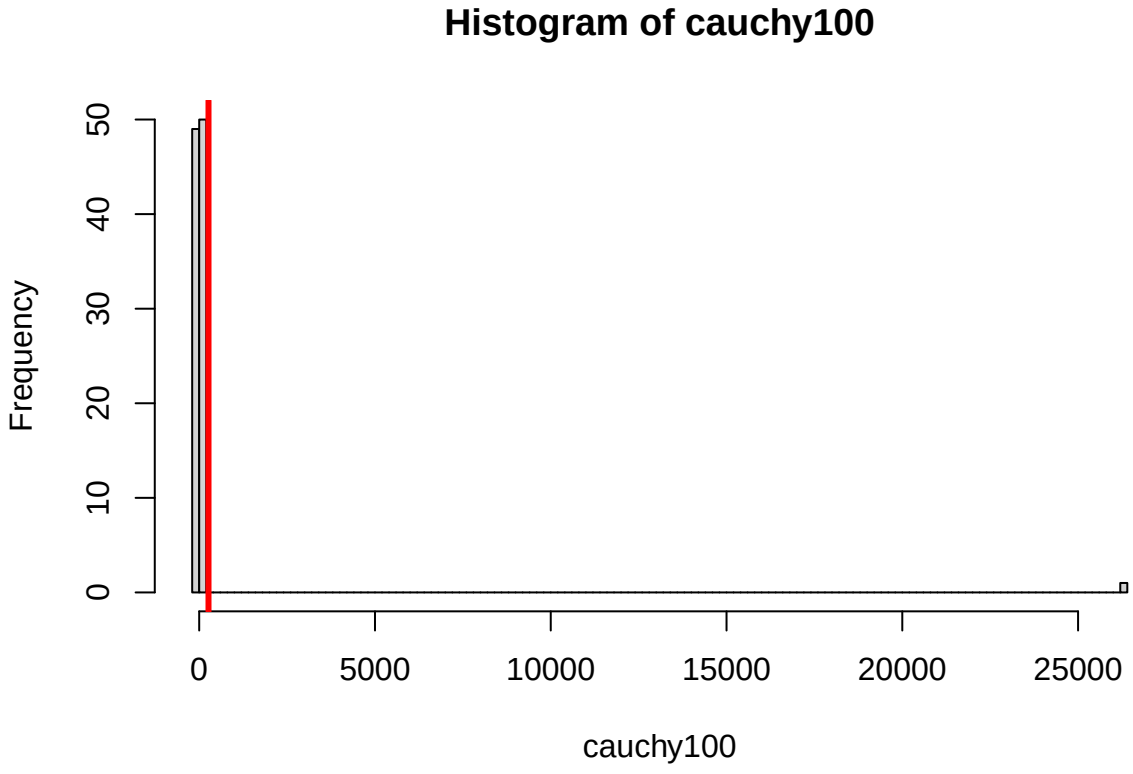
5. En déduire un estimateur de θ et evaluer la performance de cet estimateur sur les différents échantillons.

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.

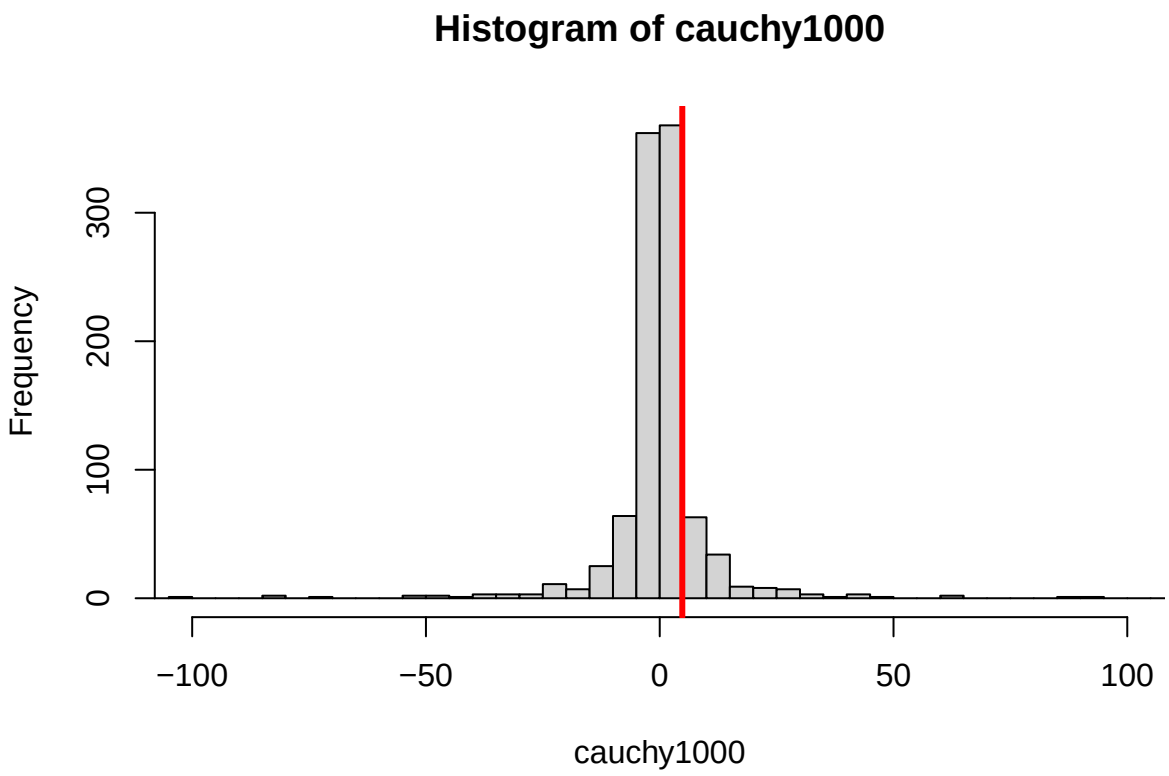
```
hist(cauchy20,breaks=20)
abline(v=mean(cauchy20),col="red",lwd=3)
```



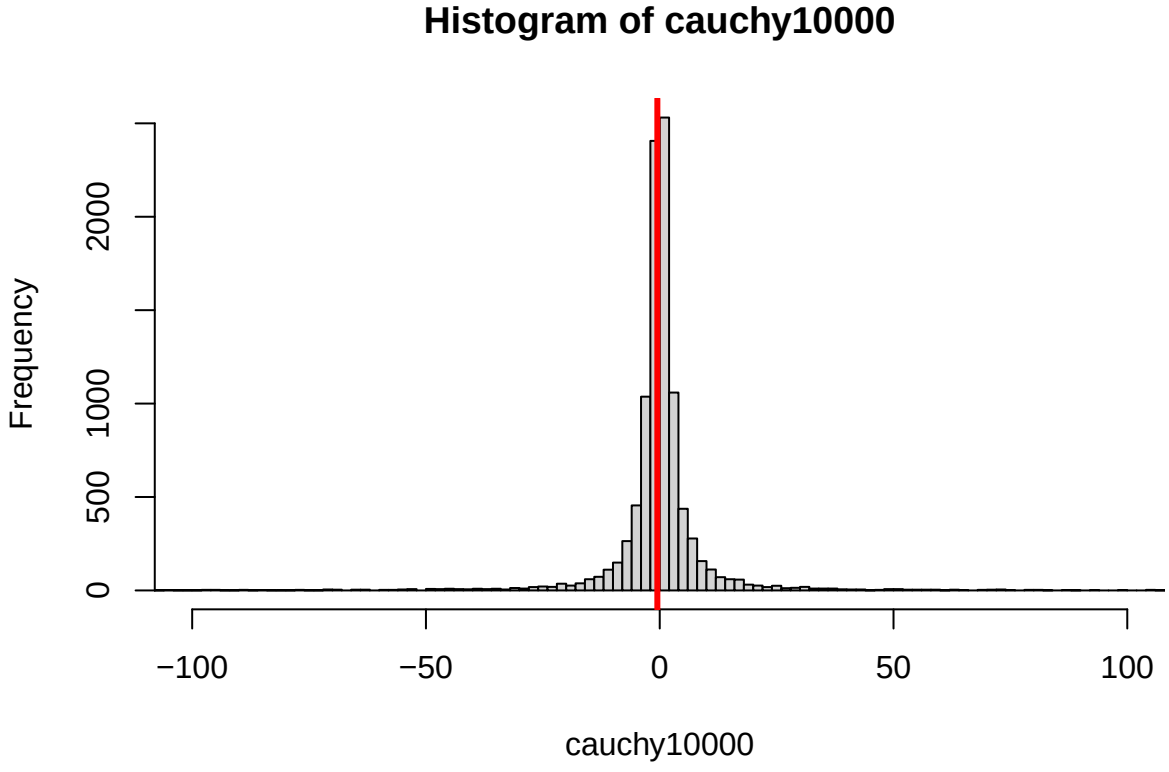

```
hist(cauchy100,breaks=100)
abline(v=mean(cauchy100),col="red",lwd=3)
```




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




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



D'après les graphiques nous pouvons remarquer que l'estimation de θ 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.