Exercice Histogramme

Rachid Sahli

23 janvier 2025

```
source("fonction_histogramme_regulier.R")
```

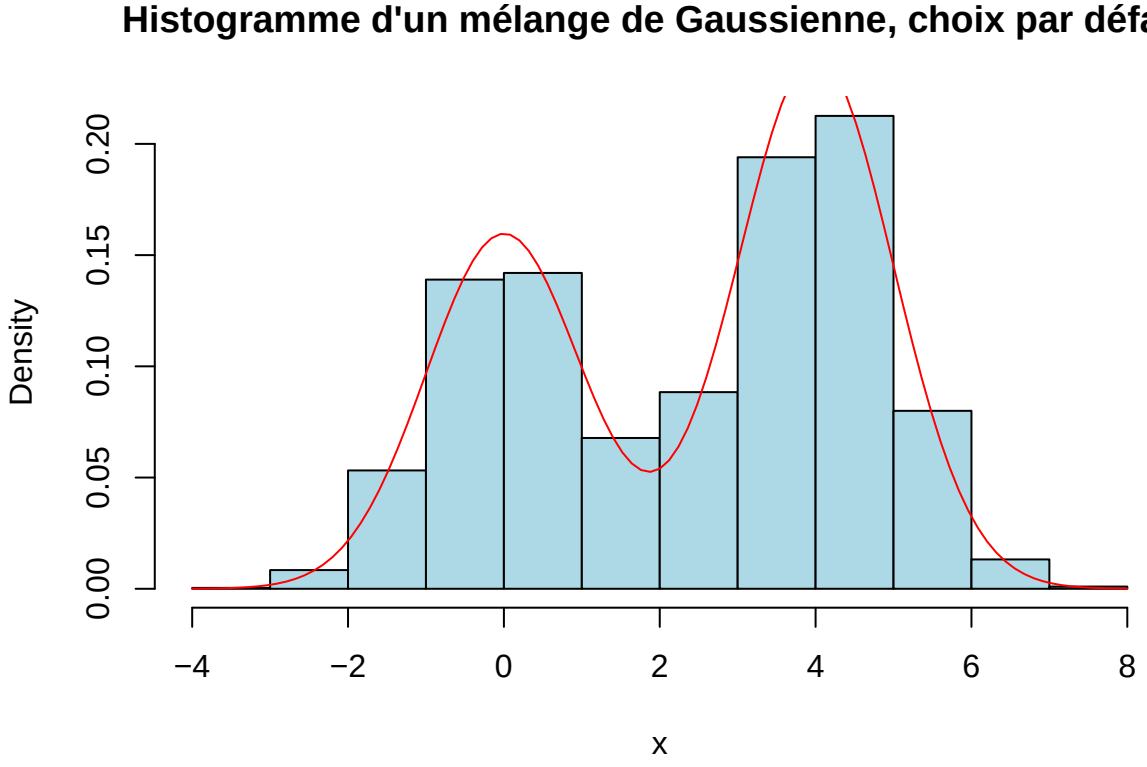
Exercice 1

Simulation d'un tirage de taille n selon un mélange de Gaussiennes, selon une loi N(0,1) et une proba 0.6 de tirer selon une loi N(0,4).

```
n <- 5000
y <- rbinom(n,1,0.4)
head(y,10)
```

```
##  [1] 0 0 1 1 1 1 0 0 0 1
```

```
x <- y*rnorm(n)+(1-y)*rnorm(n,4)
hist(x, freq = FALSE, main = "Histogramme d'un mélange de Gaussienne, choix par défaut",
     col = "lightblue")
curve(0.4*dnorm(x)+0.6*dnorm(x,4), add = TRUE, col = "red")
```

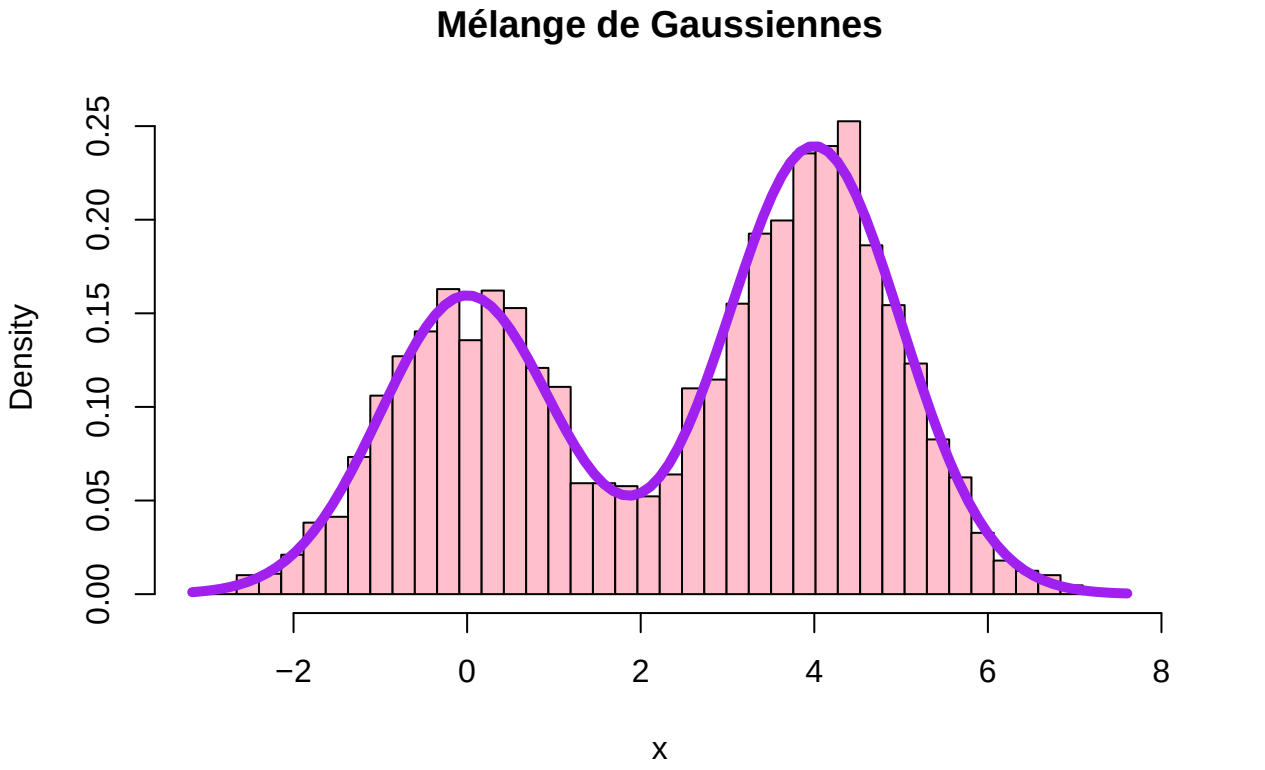


Estimation de la densité par Histogramme, en utilisant HISTSELECT, le choix par défaut de R, le nombre de breaks=[nˆ{1/3}],et la méthode de Diaconis Freedman.

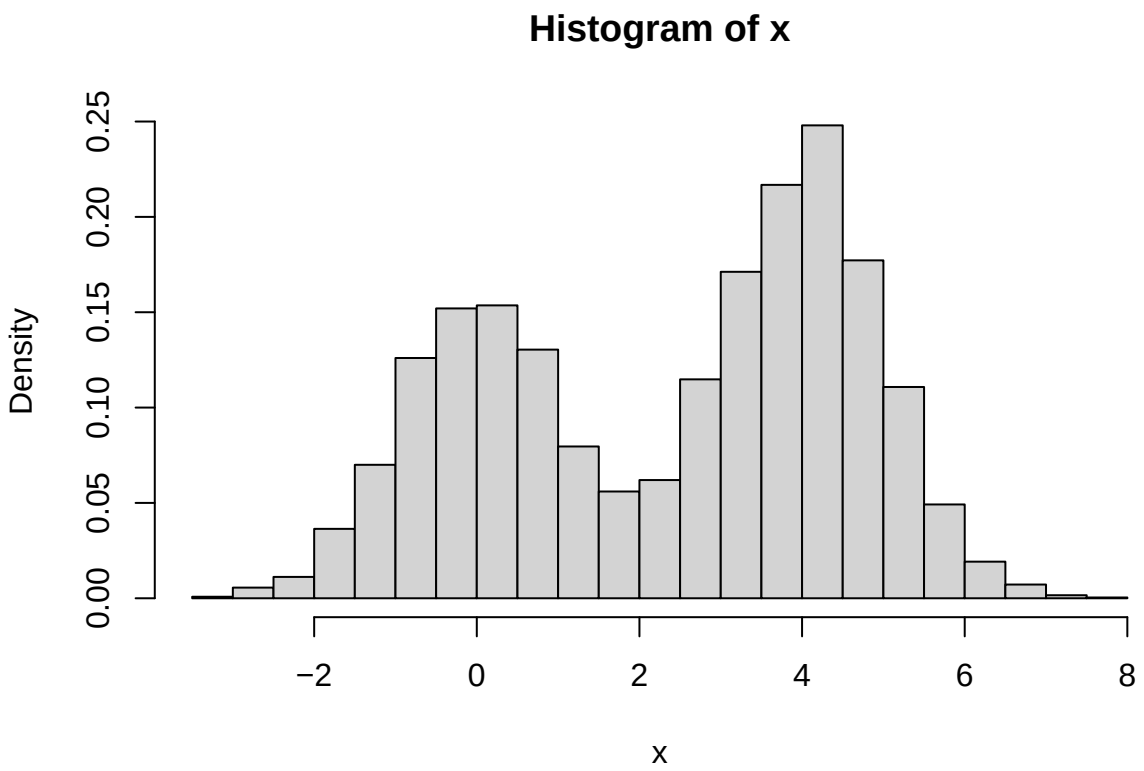
```
# Quasi optimale
HISTSELECT2(x,freq = FALSE, col = "pink", main = "Mélange de Gaussiennes")
```

```
## Number of observations in the chosen interval: 5000
## Partition size: 42
```

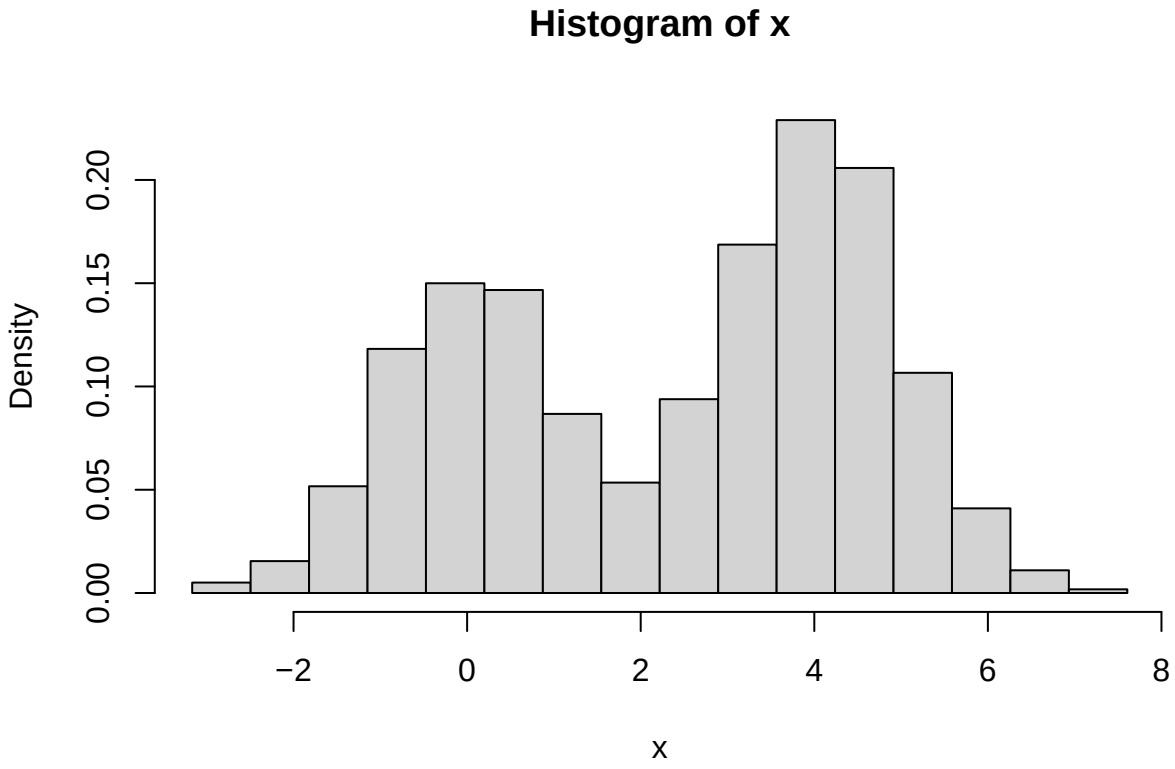
```
curve(0.4*dnorm(x)+0.6*dnorm(x,4), add = TRUE, col = "purple", lwd = 5)
```




```
hist(x, breaks = "FD", freq = FALSE)
```




```
bins <- seq(min(x),max(x), length=round(n^(1/3)))
hist(x, breaks = bins, freq = FALSE)
```

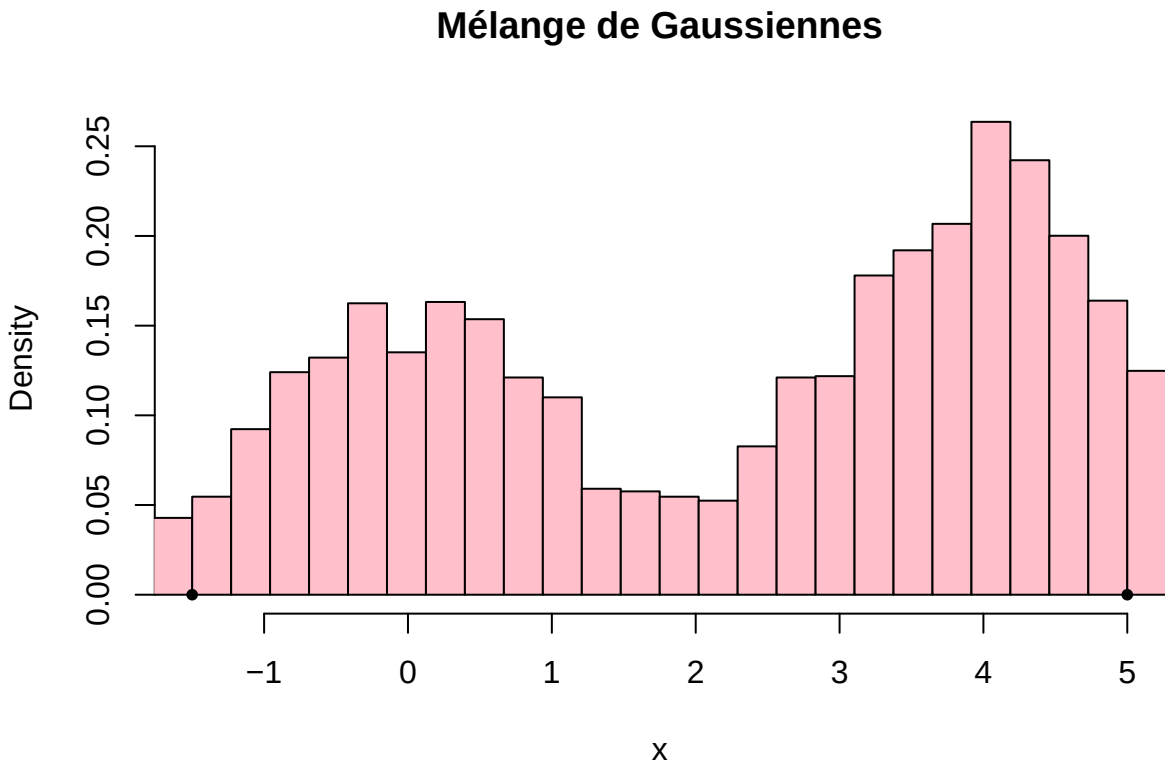


Traçage pour chaque méthode de la vraie densité et l'histogramme sur un même graphique.

Estimation de la densité sur l'intervalle [-1.5, 5] avec HISTSELECT

```
HISTSELECT2(x, freq = FALSE, col = "pink", main = "Mélange de Gaussiennes",-1.5,5)
```

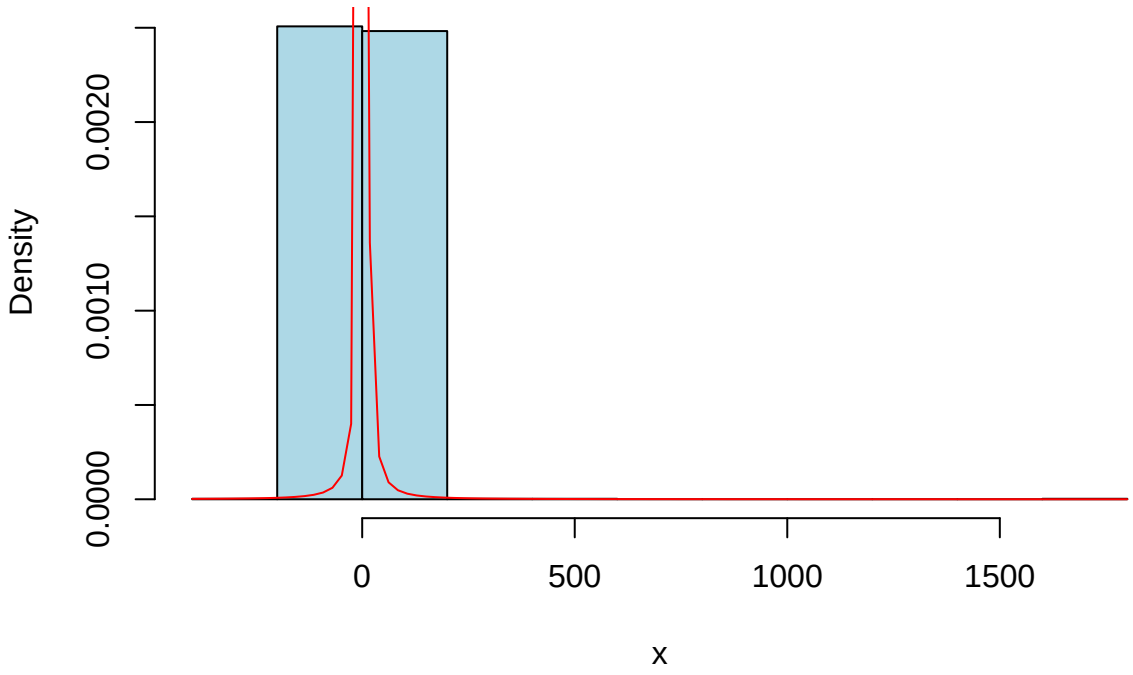
```
## Number of observations in the chosen interval: 4394
## Partition size: 24
```



Exercice 2 :

Simulation d'un tirage de taille n selon une loi de Cauchy.

```
n <- 2000
x <- rcauchy(n)
hist(x, freq = FALSE, main = "Histogramme d'une loi de cauchy, chaoix par defaut",
     col = "lightblue")
curve(0.4*dcauchy(x)+0.6*dcauchy(x,4), add = TRUE, col = "red")
```

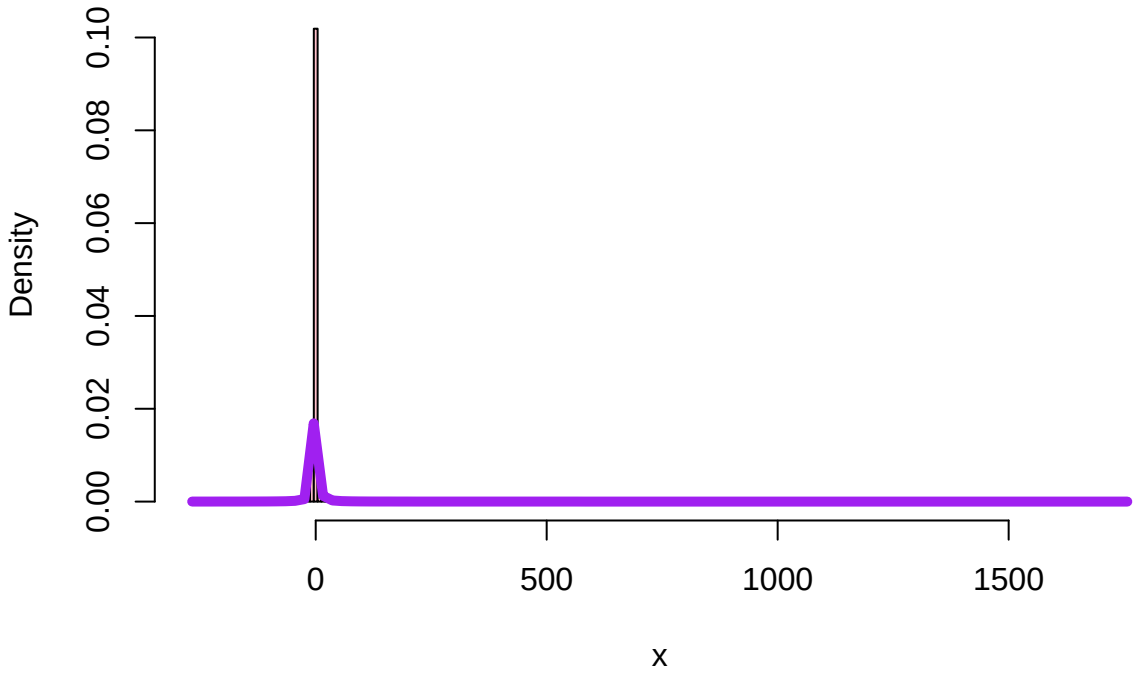


Estimation de la densité par Histogramme, en utilisant HISTSELECT, le choix par défaut de R,le nombre de breaks=[n^{1/3}],et la méthode de Diaconis Freedman.

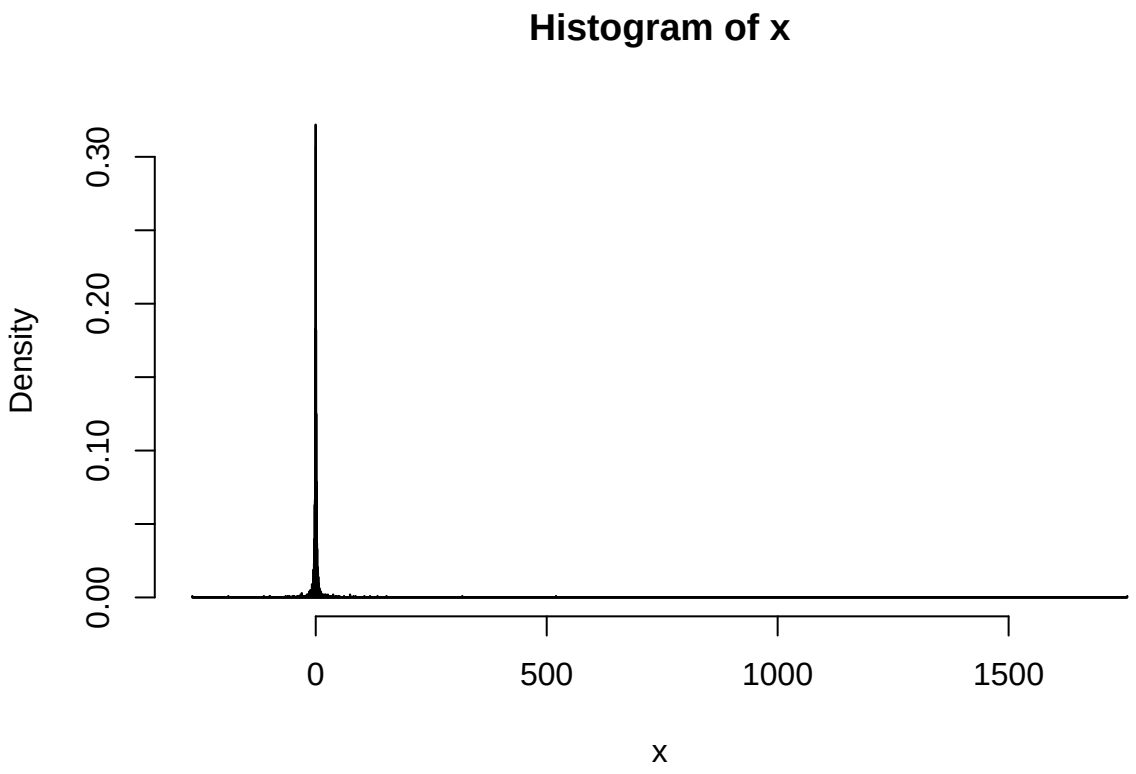
```
HISTSELECT2(x,freq = FALSE, col = "pink", main = "")
```

```
## Number of observations in the chosen interval: 2000
## Partition size: 246
```

```
curve(dcauchy(x), add = TRUE, col = "purple", lwd = 5)
```




```
hist(x, breaks = "FD", freq = FALSE)
```

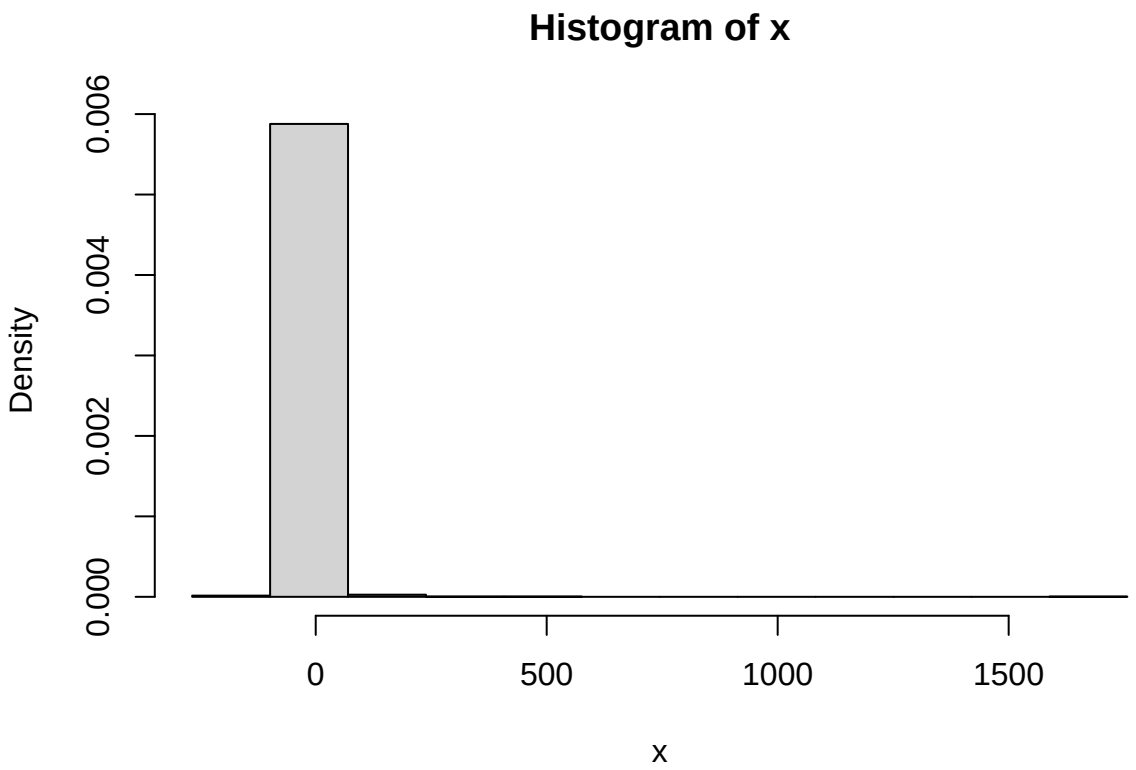



```
bins <- seq(min(x),max(x), length=round(n^(1/3)))
```

Ca ne marche pas car l'histogramme est fait pour estimer des intervalles de petite taille.

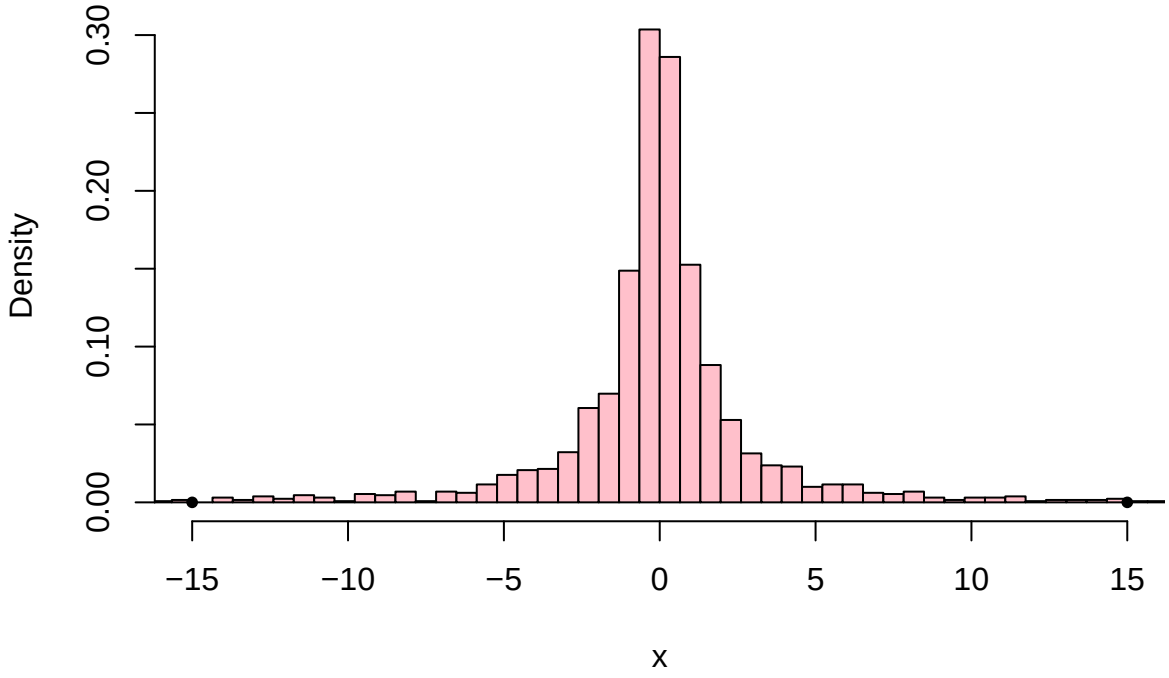
Proposer une solution à l'aide de HISTSELECT

```
hist(x, breaks = bins, freq = FALSE)
```




```
HISTSELECT2(x, freq = FALSE, col = "pink", main = "",-15,15)
```

```
## Number of observations in the chosen interval: 1914
## Partition size: 46
```

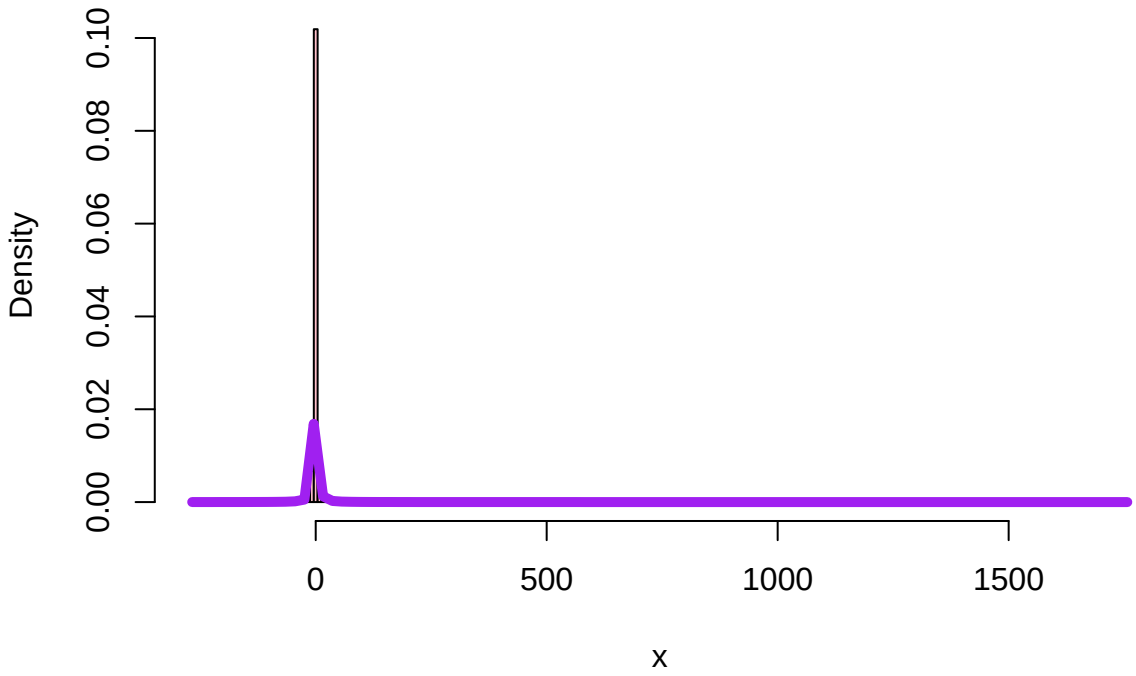


Quasi optimale

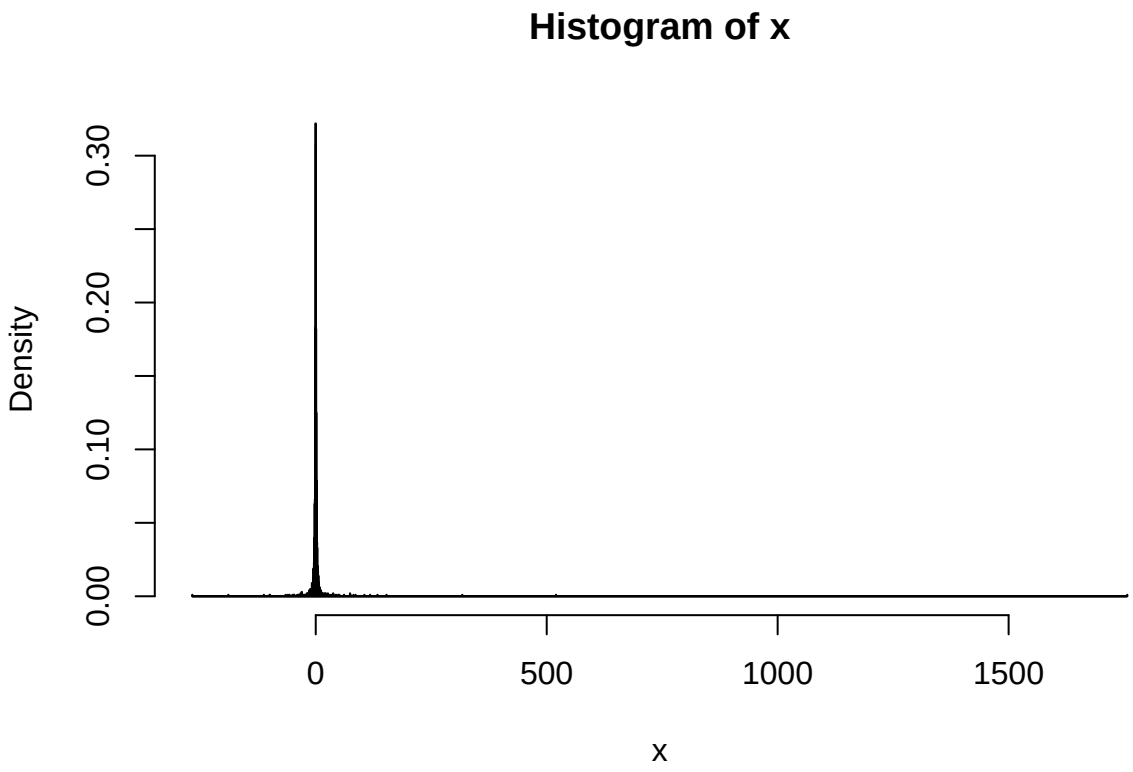
```
HISTSELECT2(x,freq = FALSE, col = "pink", main = "")
```

```
## Number of observations in the chosen interval: 2000
## Partition size: 246
```

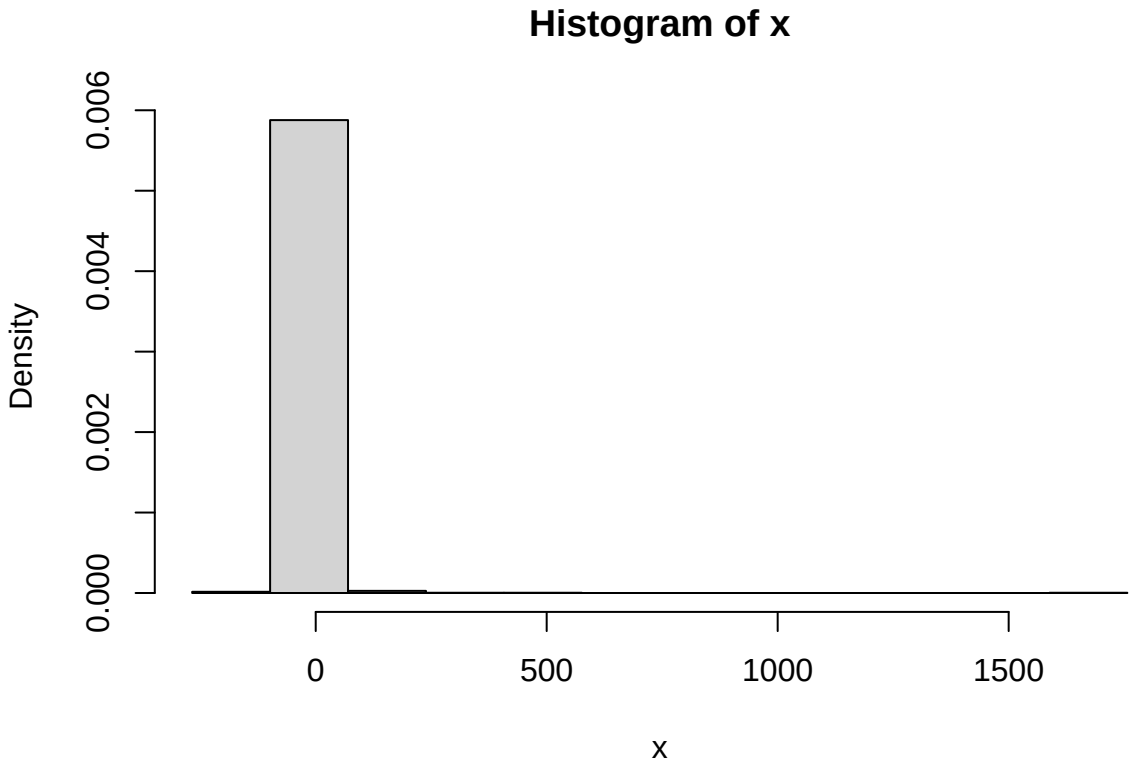
```
curve(dcauchy(x), add = TRUE, col = "purple", lwd = 5)
```




```
hist(x, breaks = "FD", freq = FALSE)
```



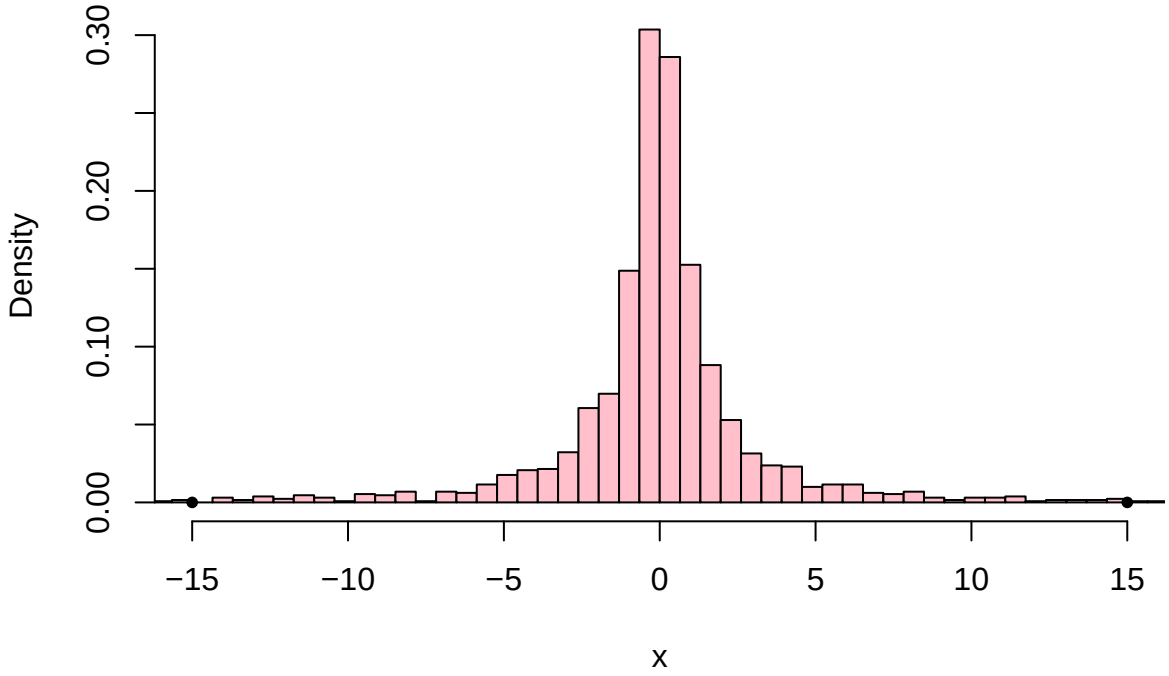

```
bins <- seq(min(x),max(x), length=round(n^(1/3)))
hist(x, breaks = bins, freq = FALSE)
```




```
HISTSELECT2(x, freq = FALSE, col = "pink", main = "",-15,15)
```

```
## Number of observations in the chosen interval: 1914
```

```
## Partition size: 46
```



Visualisation de l'écart à f (inconnu si f est inconnue) :

Loi beta

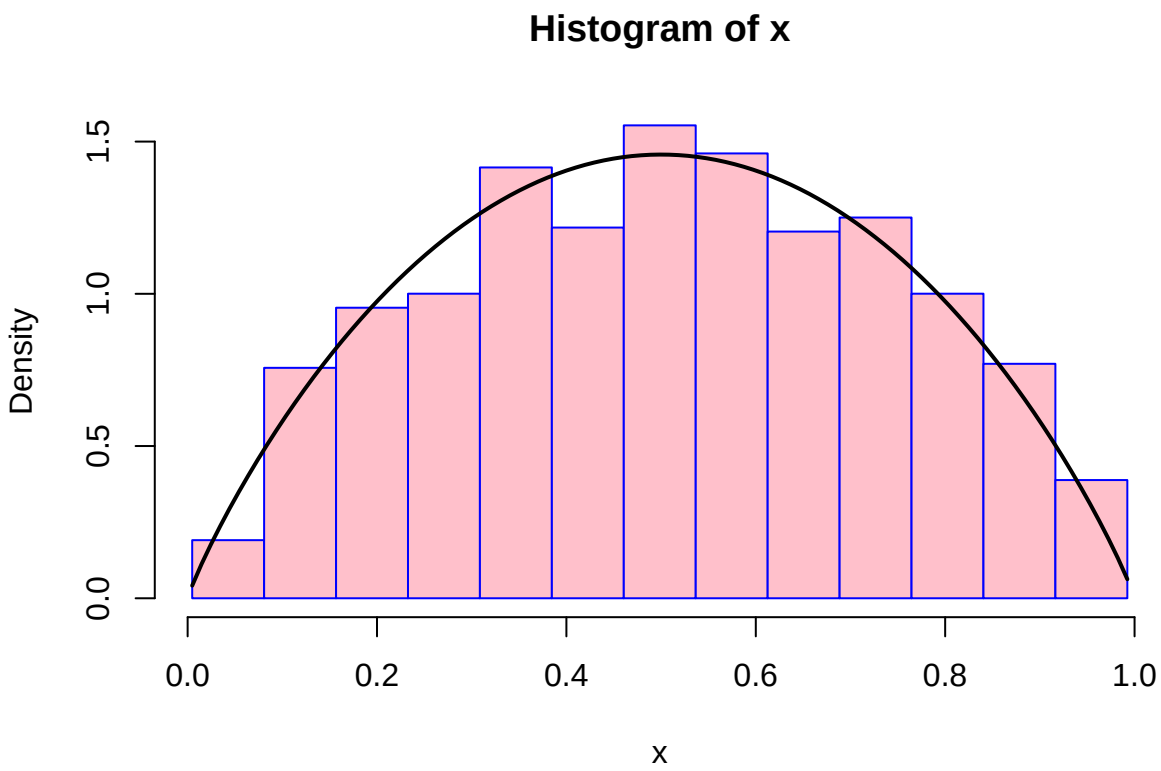
```
n2<-10000000
x<-rbeta(n2,1.9,1.9)
mean(dbeta(x, 1.9,1.9))#1.176
```

```
## [1] 1.175836
```

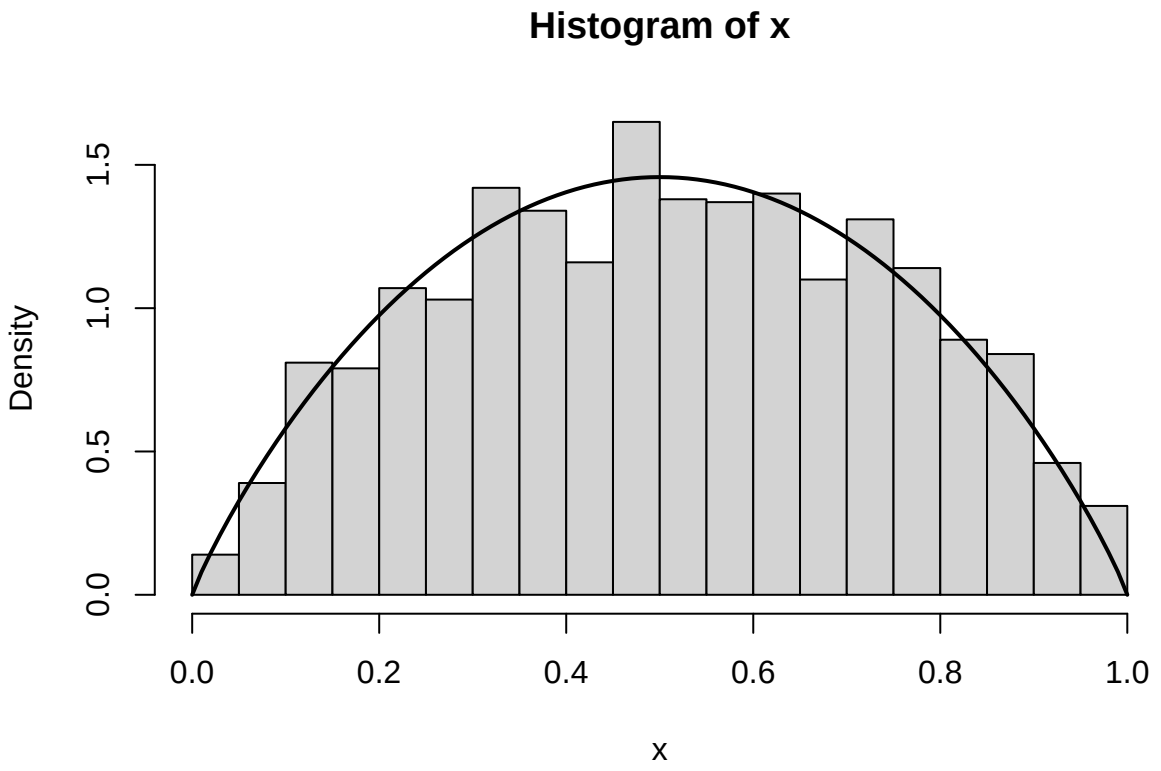
```
n<-2000
x<-rbeta(n, 1.9, 1.9)
HISTSELECT2(x,col="pink",border="blue", freq=FALSE)
```

```
## Number of observations in the chosen interval: 2000
## Partition size: 13
```

```
curve(dbeta(x, 1.9, 1.9), add=TRUE, lwd=2)
```




```
hist(x,freq=FALSE, breaks="FD")
curve(dbeta(x, 1.9, 1.9), add=TRUE, lwd=2)
```



Exemple sur jeu de données recensement

```
# Import data ----

data <- read.table("/Users/rs777/Documents/Statistique-non-parametrique/data/Recensement.txt",
                   header = TRUE)
```

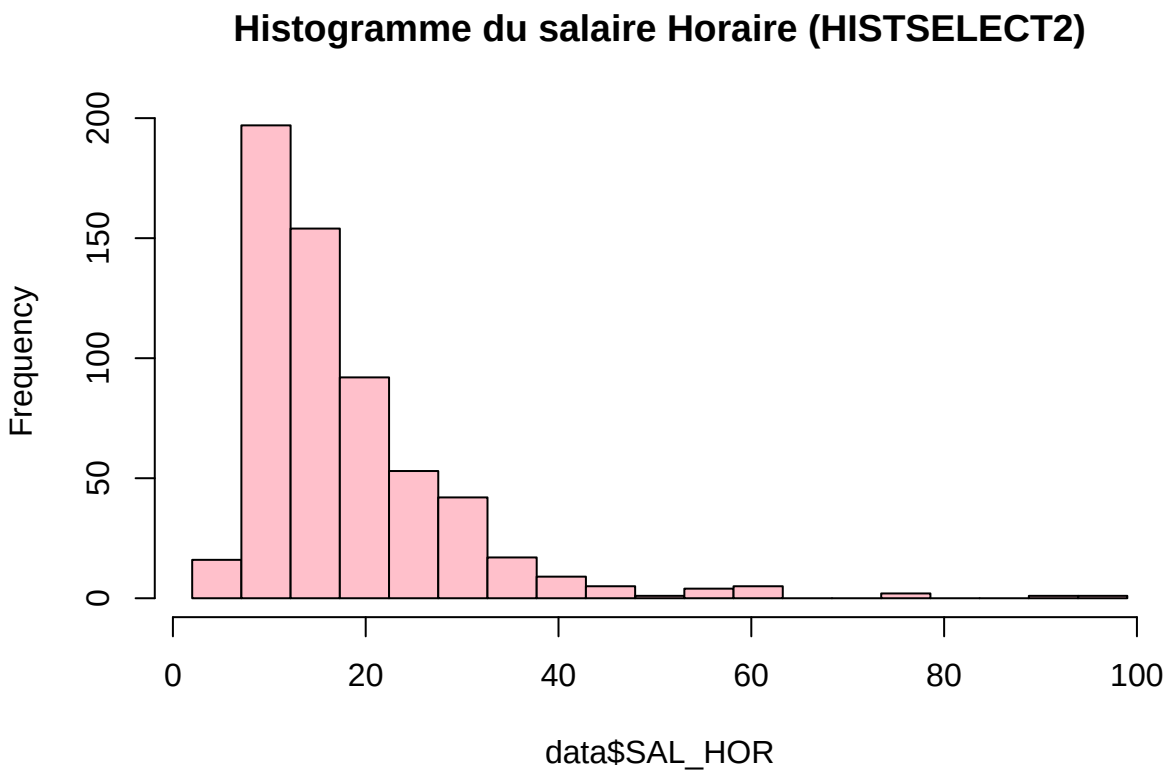
```
# Résumé statistique des variables
summary(data)
```

```
##       AGE            SEXE              REGION           STAT_MARI
##  Min.   :16.00   Length:599         Length:599         Length:599
##  1st Qu.:29.00   Class :character   Class :character   Class :character
##  Median :42.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :41.85
##  3rd Qu.:53.50
##  Max.   :80.00
##     SAL_HOR       SYNDICAT           CATEGORIE        NIV_ETUDES
##  Min.   : 2.0   Length:599         Min.   : 1.000   Min.   :32.00
##  1st Qu.:10.5   Class :character   1st Qu.: 2.000   1st Qu.:39.00
##  Median :15.0   Mode  :character   Median : 3.000   Median :40.00
##  Mean   :17.9                      Mean   : 4.387   Mean   :40.36
##  3rd Qu.:22.0                      3rd Qu.: 6.000   3rd Qu.:43.00
##  Max.   :99.0                      Max.   :10.000   Max.   :46.00
##     NB_PERS          NB_ENF         REV_FOYER
##  Min.   : 1.00   Min.   :0.0000   Min.   : 1.00
##  1st Qu.: 2.00   1st Qu.:0.0000   1st Qu.:10.00
##  Median : 3.00   Median :0.0000   Median :12.00
##  Mean   : 3.11   Mean   :0.5326   Mean   :11.57
##  3rd Qu.: 4.00   3rd Qu.:1.0000   3rd Qu.:14.00
##  Max.   :13.00   Max.   :6.0000   Max.   :16.00
```

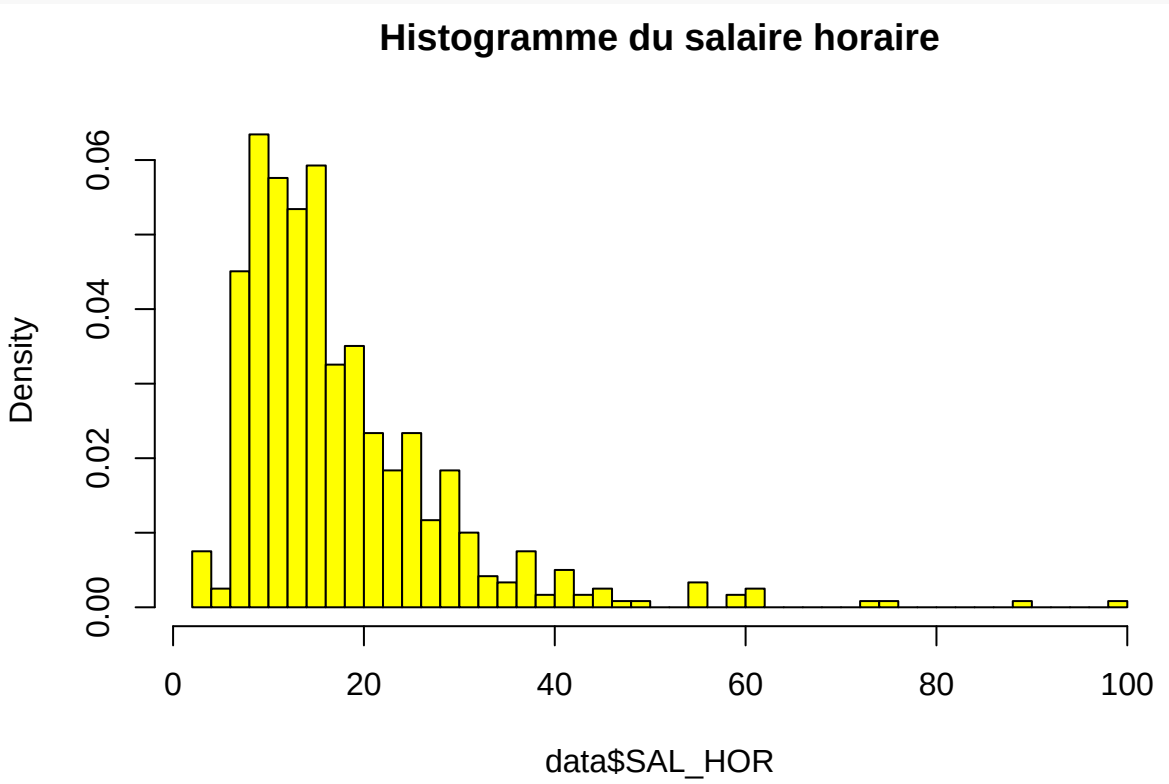
Utilisation d'HISTSELECT et de la méthode de Diaconis Freedman pour estimer la densité du Salaire Horaire et de l'Age.

```
HISTSELECT2(data$SAL_HOR, col = "pink", main = "Histogramme du salaire Horaire (HISTSELECT2)")
```

```
## Number of observations in the chosen interval: 599
## Partition size: 19
```



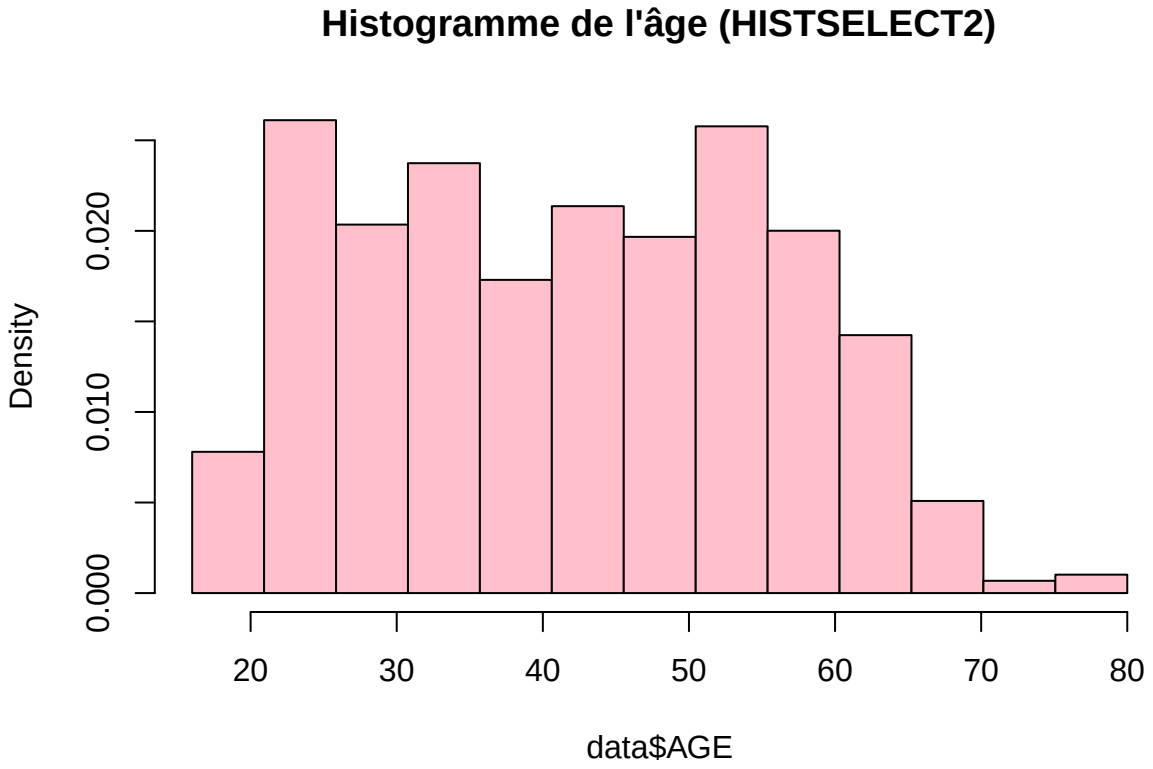

```
hist(data$SAL_HOR, col = "yellow", main = "Histogramme du salaire horaire", breaks = "FD",
     freq = FALSE)
```



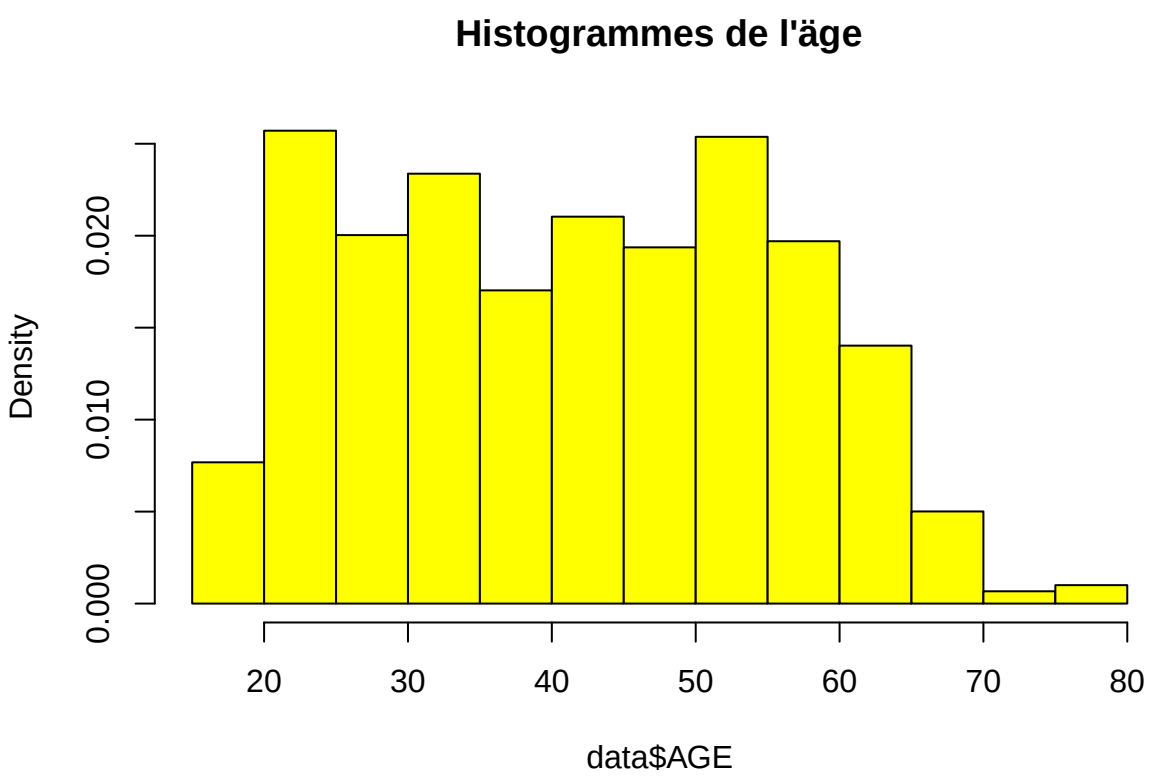

```
HISTSELECT2(data$AGE, col = "pink", main = "Histogramme de l'âge (HISTSELECT2)", freq = FALSE, nmax = 25
```

```
## Number of observations in the chosen interval: 599
```

```
## Partition size: 13
```




```
hist(data$AGE, col = "yellow", main = "Histogrammes de l'äge", breaks = "FD", freq = FALSE)
```



```r
head(data$AGE)
```

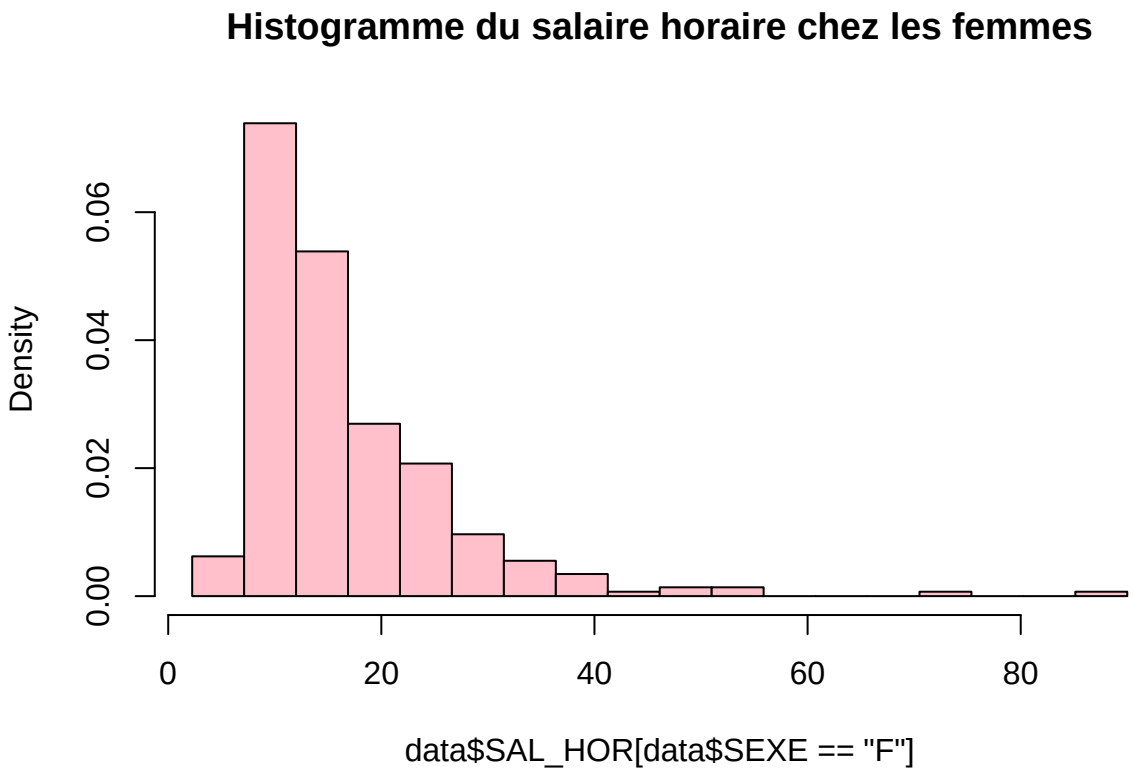
```
## [1] 58 40 29 59 51 19
```

L'histogramme généré avec le choix de classe automatique (HISTSELECT2) permet d'ajuster les intervalles de manière plus flexible et adaptée aux données, tandis que le choix par défaut de R utilise un nombre fixe de classes.

Utilisation d'HISTSELECT pour estimer la densité du salaire horaire chez les femmes et chez les hommes

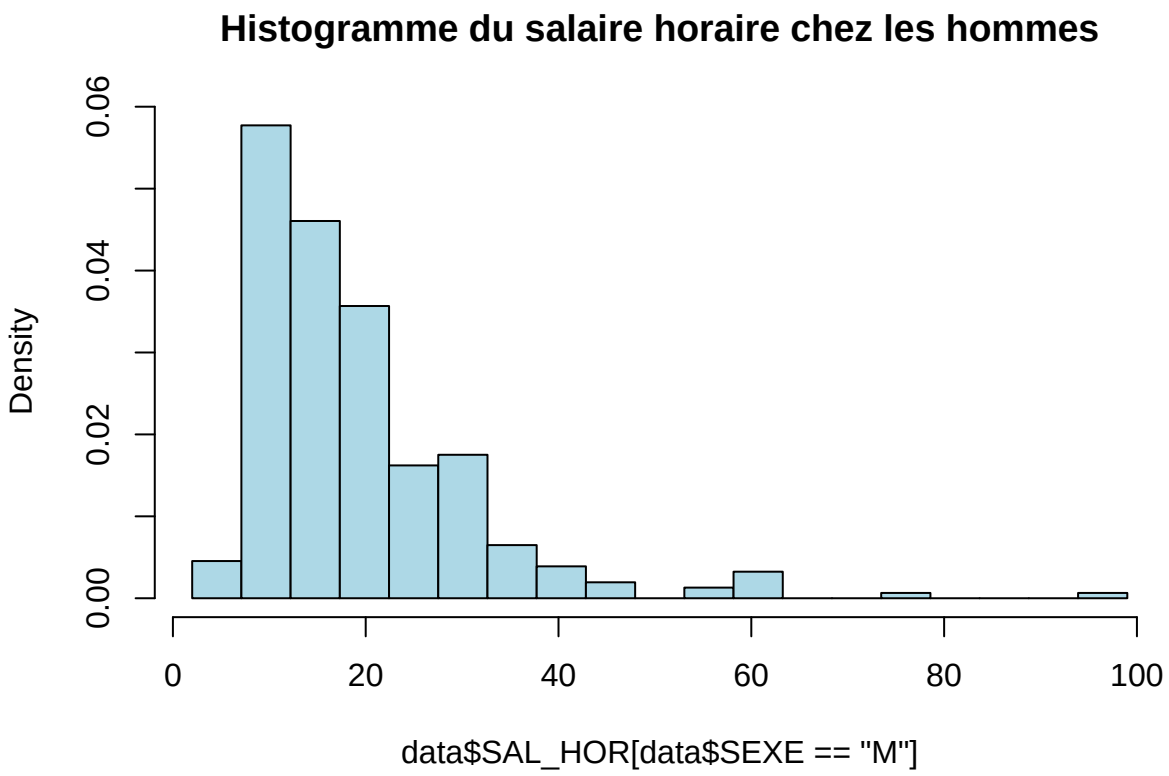
```r
HISTSELECT2(data$SAL_HOR[data$SEXE == "F"], col = "pink", main = "Histogramme du salaire horaire chez l
```

```
## Number of observations in the chosen interval: 297
## Partition size: 18
```




```r
HISTSELECT2(data$SAL_HOR[data$SEXE == "M"], col = "lightblue", main = "Histogramme du salaire horaire c
```

```
## Number of observations in the chosen interval: 302
## Partition size: 19
```


Histogramme du salaire horaire chez les hommes
Density
0.00
0.02
0.04
0.06
0
20
40
60
80
100
data$SAL_HOR[data$SEXE == "M"]