MUSÉUM NATIONAL D'HISTOIRE NATURELLE

SEMIN-R

Histogrammes

Séverine ZIRAH
Molécules de communication et adaptation des micro-organismes
FRE 3206 CNRS / USM 502 MNHN

Sémin'R – 22.01.2009

REPRESENTATIONS GRAPHIQUES SOUS R

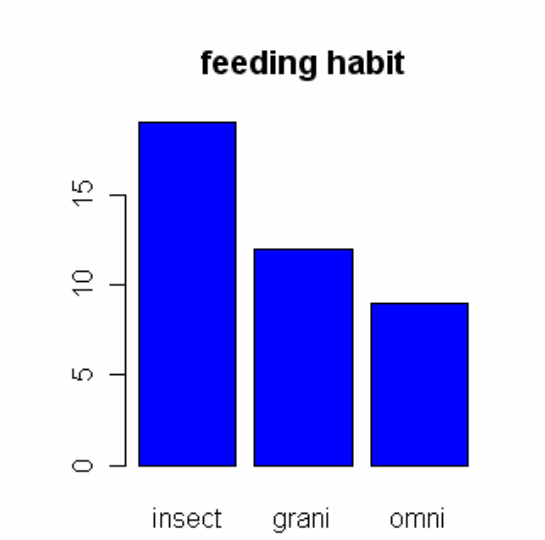
HISTOGRAMMES

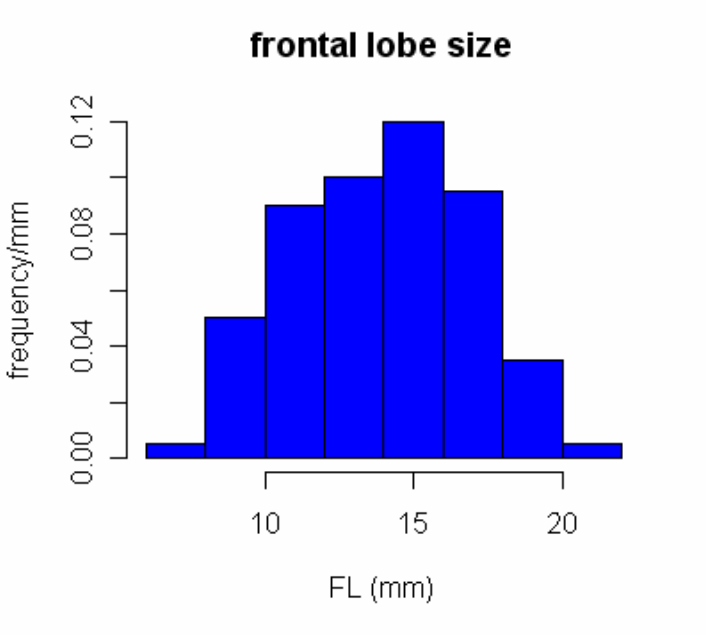
Séverine ZIRAH
Molécules de communication et adaptation des micro-organismes
FRE 3206 CNRS / USM 502 MNHN

Représentation graphique d'une distribution

Variable

- **qualitative**
- **quantitative discrète**
- **quantitative continue**

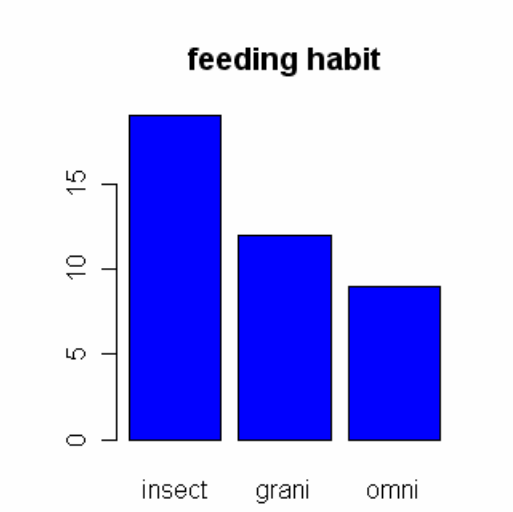


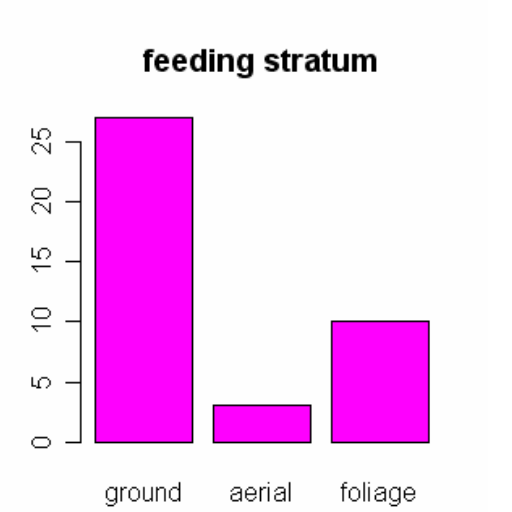


Représentation graphique d'une variable qualitative

Diagramme en bâton ➡ plot

```
> library(ade4)
Warning message:
package 'ade4' was built under R version 2.3.1
> data(aviurba)
> aviurba$traits$feed.hab
 [1] omni   omni   omni   grani  grani  grani  insect insect insect
[10] insect insect insect omni   insect insect insect insect insect
[19] omni   insect insect insect insect insect insect insect grani
[28] grani  grani  grani  grani  grani  grani  grani  grani  insect
[37] omni   omni   omni   omni
Levels: insect grani omni
> plot(aviurba$traits[,1],main="feeding habit",col="blue")
> plot(aviurba$traits[,2],main="feeding stratum",col="magenta")
```





Représentation graphique à partir d'un tableau d'effectifs/de fréquences

Tableau d'effectifs ➡ table

Classe	Effectif	Fréquence	%
1	n_1	$f_1 = \frac{n_1}{n}$	$100\times f_1$
⋮	⋮	⋮	⋮
i	n_i	$f_i = \frac{n_i}{n}$	$100\times f_i$
⋮	⋮	⋮	⋮
p	n_p	$f_p = \frac{n_p}{n}$	$100\times f_p$
TOTAL	**n**	**1**	**100**

```
> table(aviurba$traits$feed.hab)
insect  grani   omni
    19     12      9
> table(aviurba$traits$feed.strat)
 ground  aerial foliage
     27       3      10
> table(aviurba$traits$breeding)
  ground building    scrub  foliage
       6       14       12        8
> table(aviurba$traits$migratory)
resident  migrant
      23       17
```

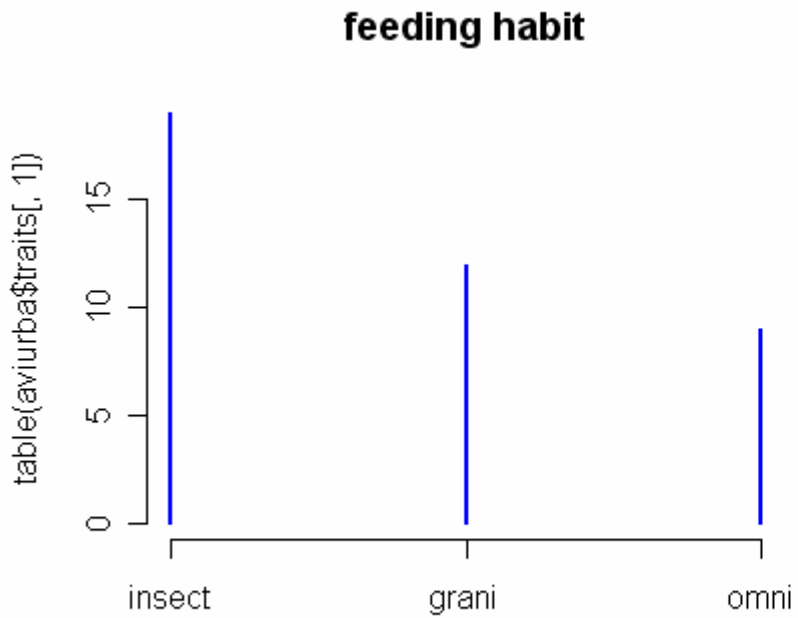
Diagramme en bâton ➡ plot

barplot

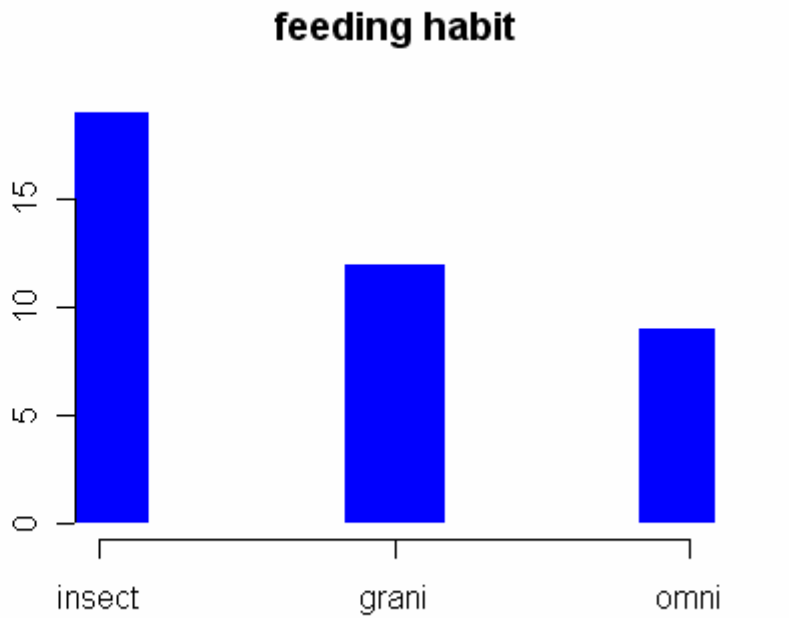
Représentation graphique à partir d'un tableau d'effectifs/de fréquences

plot

```
> plot(table(aviurba$traits[,1]),
col="blue", main="feeding habit")
```



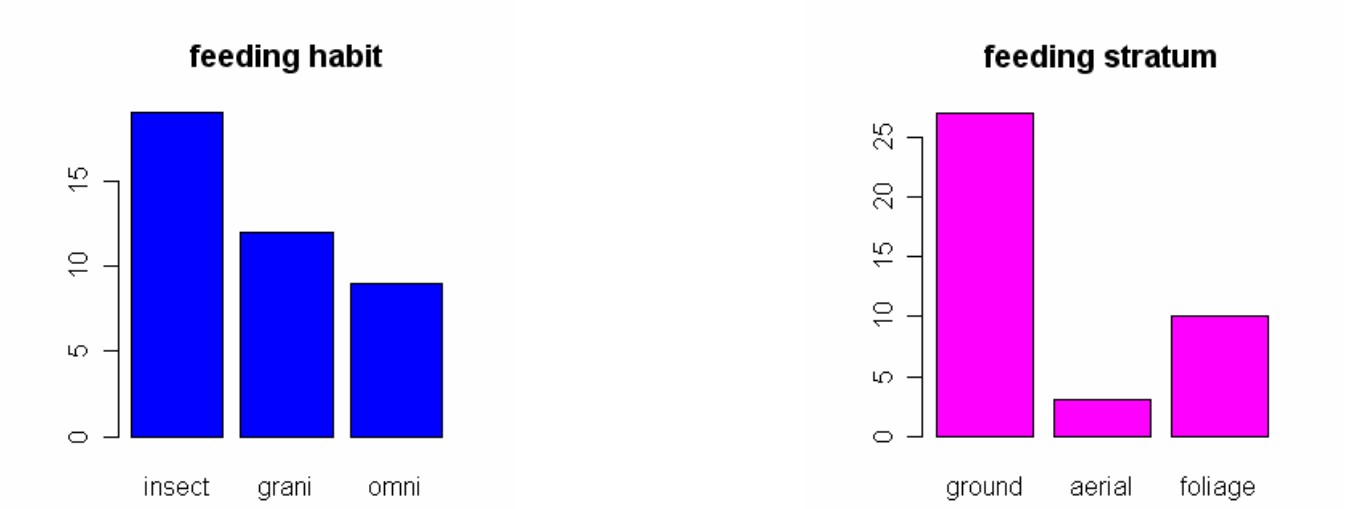

```
> plot(table(aviurba$traits[,1]),
col="blue", main="feeding habit",
lwd=50,type="h",lend="butt",ylab="")
```

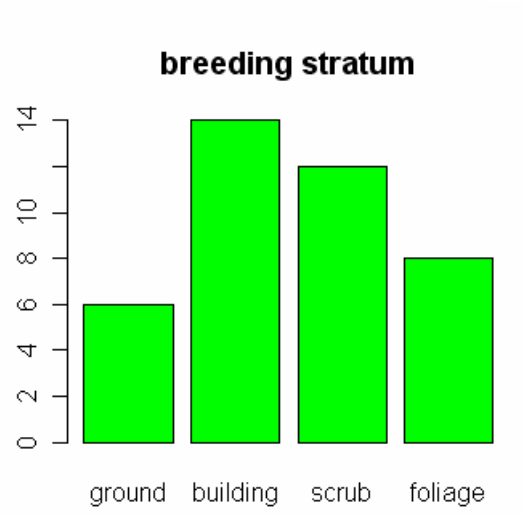


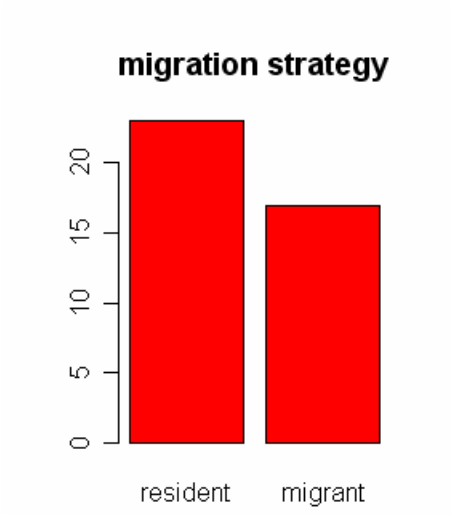
Représentation graphique à partir d'un tableau d'effectifs/de fréquences

barplot

```
> barplot(table(aviurba$traits[,1]),main="feeding habit",col="blue")
> barplot(table(aviurba$traits[,2]),main="feeding stratum",col="magenta")
> barplot(table(aviurba$traits[,3]),main="breeding stratum",col="green")
> barplot(table(aviurba$traits[,4]),main="migration stragegy",col="red")
```





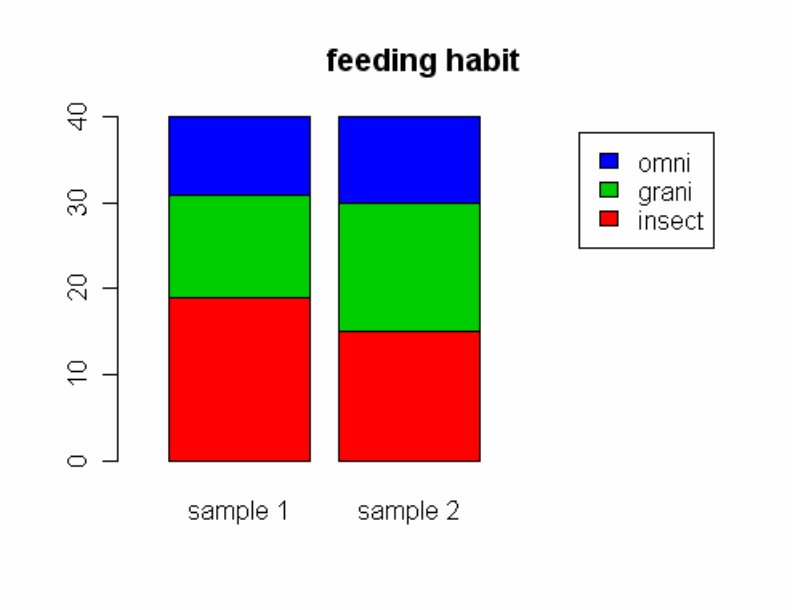


Représentation graphique d'une variable qualitative

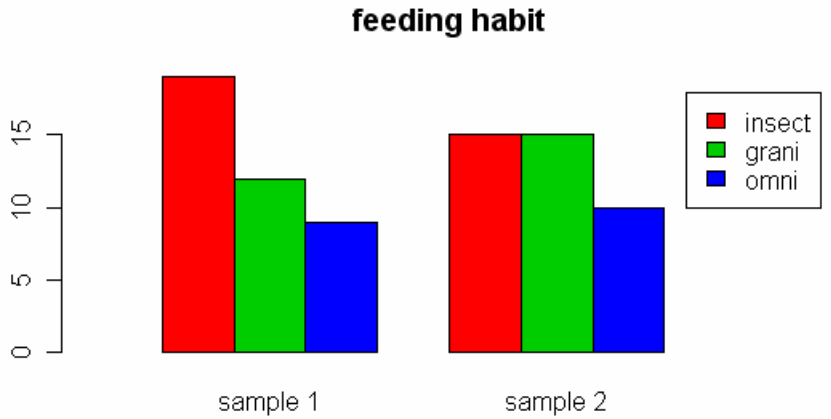
Diagrammes en bâton multiples ➡ barplot

```
> t<-cbind(table(aviurba$traits[,1]),c(15,15,10))
```

```
> barplot(t,legend.text=rownames(t),
col=c(2:4),main="feeding habit",
names=c("sample 1","sample 2"),
xlim=c(0,8),width=2)
```



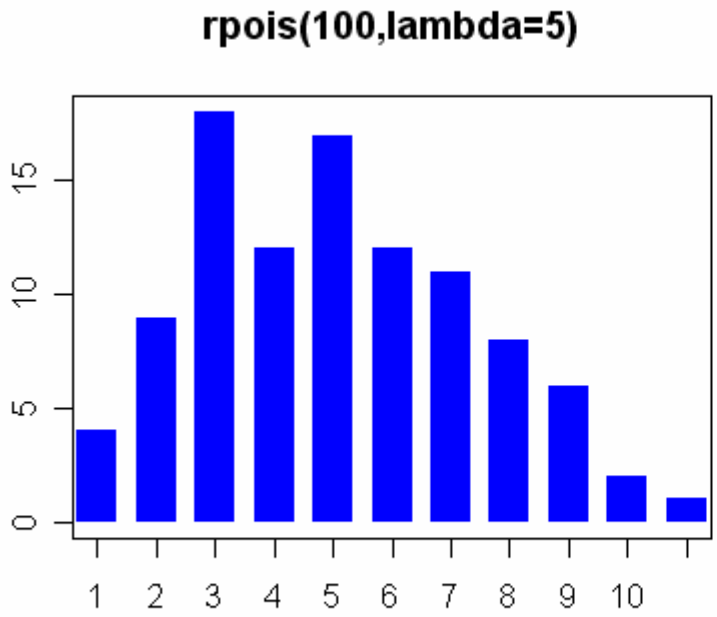

```
> barplot(t,legend.text=rownames(t),
col=c(2:4),main="feeding habit",
names=c("sample 1","sample 2"),
xlim=c(0,20),width=2,beside=TRUE)
```



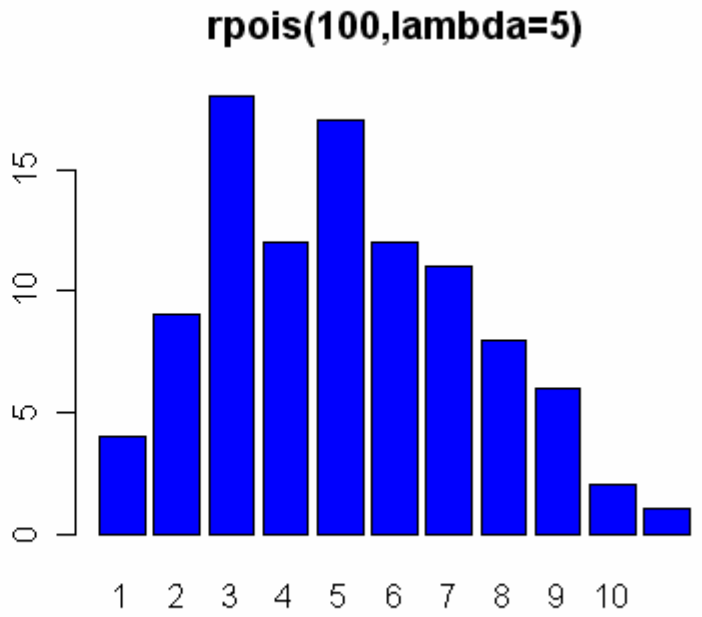
Représentation graphique d'une distribution discrète

```
> t<-table(rpois(100,5))
> t
 1  2  3  4  5  6  7  8  9 10 11
 4  9 18 12 17 12 11  8  6  2  1
> plot(t,col="blue",main="rpois(100,lambda=5)",lwd=20,type="h",lend="butt",ylab="")
> barplot(t,col="blue",main="rpois(100,lambda=5)")
```

plot



barplot



Représentation graphique d'une variable quantitative continue

Tableau d'effectifs / de fréquences

Les données sont regroupées en classes correspondant à des intervalles de valeurs.

$$\text{Valeur centrale} = \frac{\text{borne inférieure} + \text{borne supérieure}}{2}$$

Classe	Valeur (ou valeur centrale)	Effectif	Fréquence
1	V_1	n_1	$f_1 = \frac{n_1}{n}$
⋮	⋮	⋮	⋮
i	V_i	n_i	$f_i = \frac{n_i}{n}$
⋮	⋮	⋮	⋮
p	v_p	n_p	$f_p = \frac{n_p}{n}$
TOTAL		**n**	**1**

Paramètres à bien choisir :

Largeur de classe

Nombre de classes

Point de départ

Histogramme

Juxtaposition de rectangles dont les bases correspondent aux intervalles de valeurs des classes successives et dont les aires correspondent aux fréquences des classes.
Variante : Hauteurs des rectangles = effectifs par classe

Histogrammes

hist

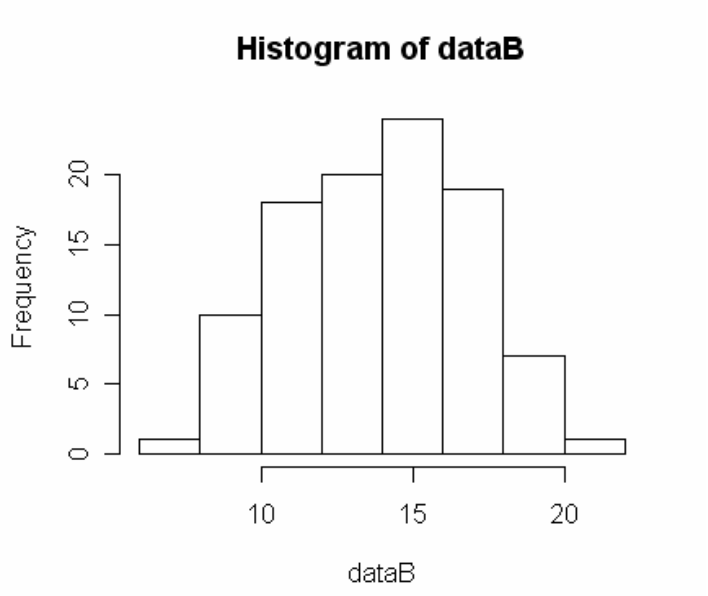
```
> help(hist)
```

```
Histograms

Description:

     The generic function 'hist' computes a histogram of the given data
     values.  If 'plot=TRUE', the resulting object of 'class
     "histogram"' is plotted by 'plot.histogram', before it is
     returned.

Usage:

     hist(x, ...)

     ## Default S3 method:
     hist(x, breaks = "Sturges", freq = NULL, probability = !freq,
          include.lowest = TRUE, right = TRUE,
          density = NULL, angle = 45, col = NULL, border = NULL,
          main = paste("Histogram of" , xname),
          xlim = range(breaks), ylim = NULL,
          xlab = xname, ylab,
          axes = TRUE, plot = TRUE, labels = FALSE,
          nclass = NULL, ...)
```

Histogrammes - exemple

Frontal lobe size

```
> library(MASS)
> data(crabs)
> str(crabs)
`data.frame':   200 obs. of  8 variables:
 $ sp   : Factor w/ 2 levels "B","O": 1 1 1 1 1 1 1 1 1 1 ...
 $ sex  : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
 $ index: int  1 2 3 4 5 6 7 8 9 10 ...
 $ FL   : num  8.1 8.8 9.2 9.6 9.8 10.8 11.1 11.6 11.8 11.8 ...
 $ RW   : num  6.7 7.7 7.8 7.9 8 9 9.9 9.1 9.6 10.5 ...
 $ CL   : num  16.1 18.1 19 20.1 20.3 23 23.8 24.5 24.2 25.2 ...
 $ CW   : num  19 20.8 22.4 23.1 23 26.5 27.1 28.4 27.8 29.3 ...
 $ BD   : num  7 7.4 7.7 8.2 8.2 9.8 9.8 10.4 9.7 10.3 ...
> dataB<-crabs[which(crabs$sp=="B"),4]
> dataO<-crabs[which(crabs$sp=="O"),4]
> hist(dataB)
```



Histogramme sans « plot »

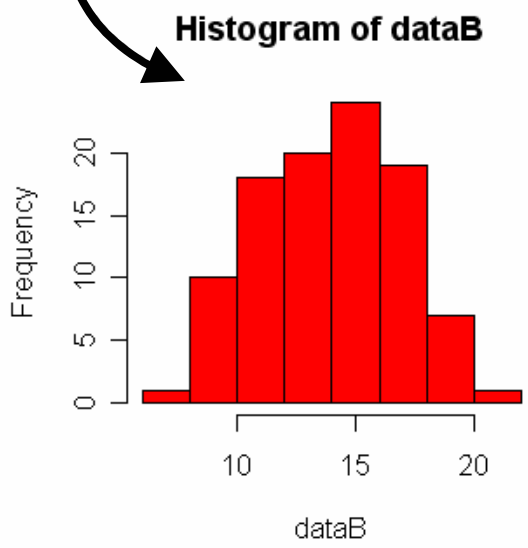
```
> hist(dataB,plot=FALSE)
$breaks
[1]  6  8 10 12 14 16 18 20 22

$counts
[1]  1 10 18 20 24 19  7  1

$intensities
[1] 0.004999999 0.050000000 0.090000000 0.100000000 0.120000000 0.095000000
[7] 0.035000000 0.005000000

$density
[1] 0.004999999 0.050000000 0.090000000 0.100000000 0.120000000 0.095000000
[7] 0.035000000 0.005000000

$mids
[1]  7  9 11 13 15 17 19 21

$xname
[1] "dataB"

$equidist
[1] TRUE

attr(,"class")
[1] "histogram"
```

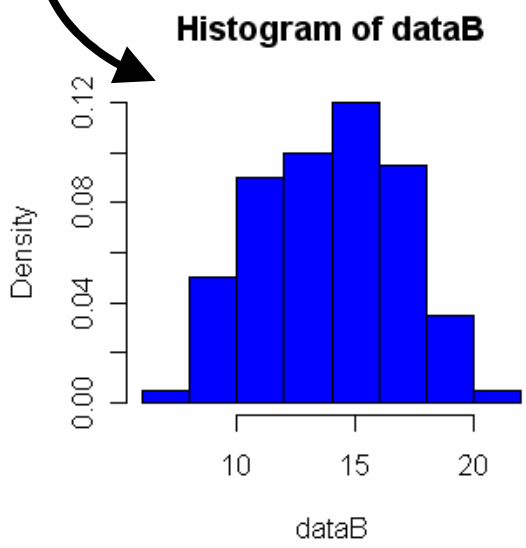
Histogrammes et tableaux d'effectifs/de fréquences

Frontal lobe size (mm)	h$counts Effectif	Fréquence	h$density Fréquence par mm
[6-8]	1	0.01	0.005
]8-10]	10	0.10	0.050
]10-12]	18	0.18	0.090
]12-14]	20	0.20	0.100
]14-16]	24	0.24	0.120
]16-18]	19	0.19	0.095
]18-20]	7	0.07	0.035
]20-22]	1	0.01	0.005
TOTAL	**100**	**1**	**0.5**

← Pour calculer la hauteur des rectangles, il suffit de diviser la fréquence par la longueur des classes.

right = "TRUE" : intervalles **]a,b]**
include.lowest = "TRUE" : premier intervalle **[a,b]**





Histogrammes et nombre de classes

breaks: one of:

* a vector giving the breakpoints between histogram cells,

* a single number giving the number of cells for the histogram,

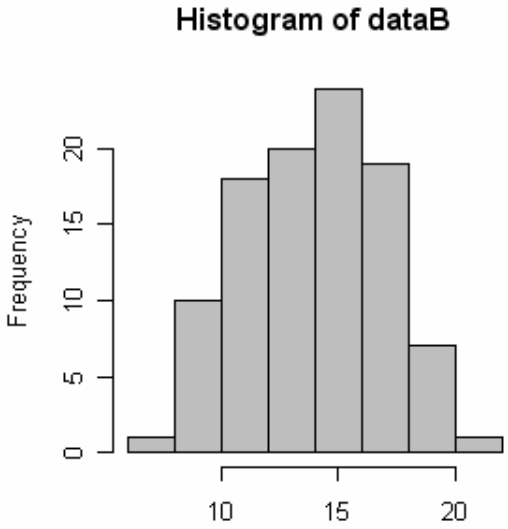
* a character string naming an algorithm to compute the number of cells (see Details),

* a function to compute the number of cells.

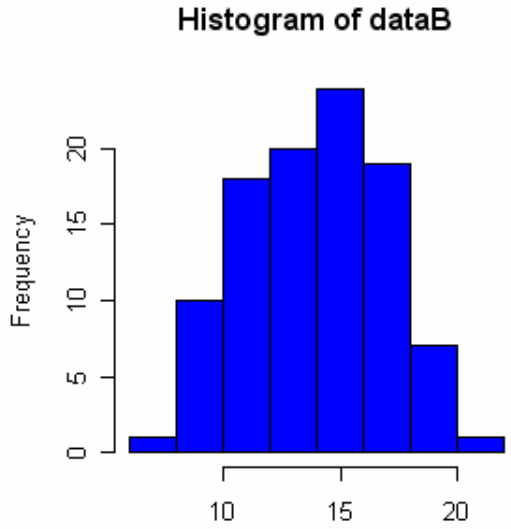
In the last three cases the number is a suggestion only.

```
> sort(dataB)
 [1]  7.2  8.1  8.8  9.0  9.1  9.1  9.2  9.5  9.6  9.8  9.8 10.1 10.3 10.4
[15] 10.8 10.8 11.0 11.1 11.2 11.5 11.6 11.6 11.6 11.7 11.8 11.8 11.9 12.0
[29] 12.0 12.2 12.3 12.6 12.6 12.8 12.8 12.8 12.8 12.8 12.9 13.0 13.1 13.1
[43] 13.1 13.2 13.3 13.4 13.7 13.9 13.9 14.3 14.6 14.7 14.9 15.0 15.0 15.0
[57] 15.0 15.0 15.1 15.1 15.1 15.2 15.2 15.3 15.4 15.4 15.5 15.6 15.6 15.7
[71] 15.7 15.8 15.9 16.1 16.1 16.2 16.2 16.3 16.4 16.4 16.6 16.7 16.8 16.9
[85] 17.1 17.1 17.2 17.4 17.5 17.7 17.9 18.0 18.8 19.2 19.3 19.3 19.7 19.8
[99] 19.8 21.3
```

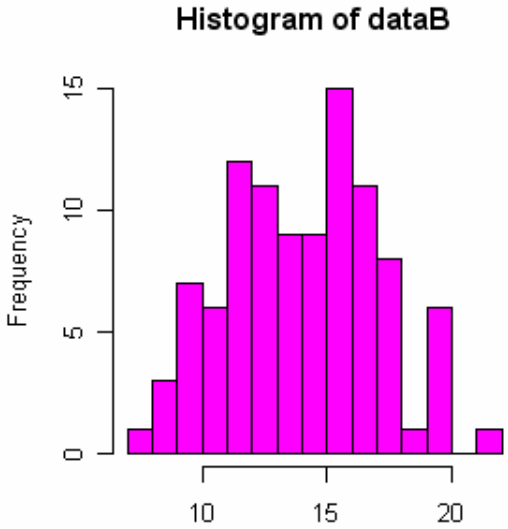
Histogrammes et nombre de classes



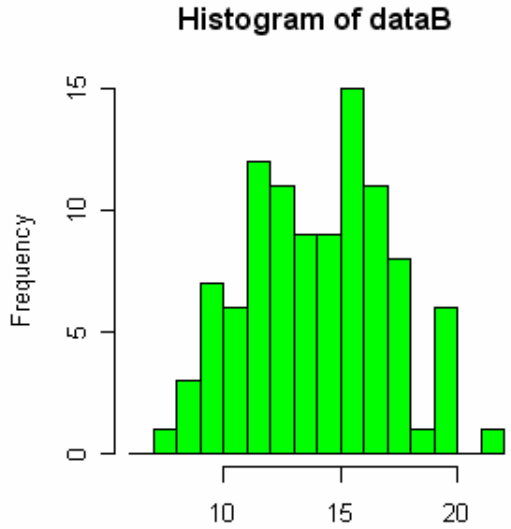

```
> hist(dataB,col="grey")
```




```
> hist(dataB,col="blue",breaks=seq(6,22,2))
```




```
> hist(dataB,col="magenta",breaks=20)
```




```
> hist(dataB,col="green",breaks=seq(6,22,1))
```

Histogrammes et nombre de classes

Algorithmes de calcul du nombre de classes

```
> help(nclass.Sturges)
```

Formule de Sturges: **la taille des classes est basée sur l'effectif de l'échantillon (nombre d'observations).**

$$\text{Nombre de classes} = 1 + 3{,}3 \log_{10}(n)$$

```
> nclass.Sturges
function (x)
ceiling(log2(length(x)) + 1)
```

Formule de Scott: **pour une distribution normale. Basée sur une estimation de la variance.**

```
> nclass.scott
function (x)
{
    h <- 3.5 * sqrt(var(x)) * length(x)^(-1/3)
    ceiling(diff(range(x))/h)
}
```

Formule de Freedman-Diaconis (FD): **basé sur l'espace inter-quartile.**

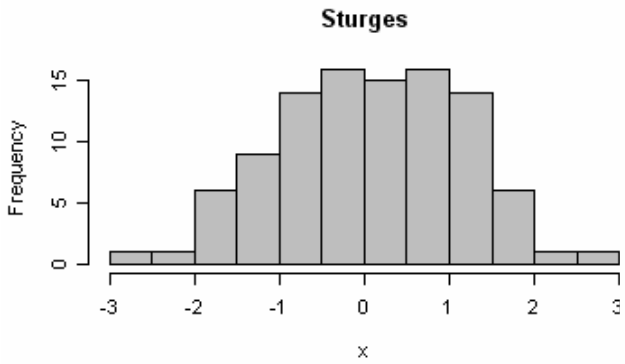
```
> nclass.FD
function (x)
{
    r <- as.vector(quantile(x, c(0.25, 0.75)))
    h <- 2 * (r[2] - r[1]) * length(x)^(-1/3)
    ceiling(diff(range(x))/h)
}
```

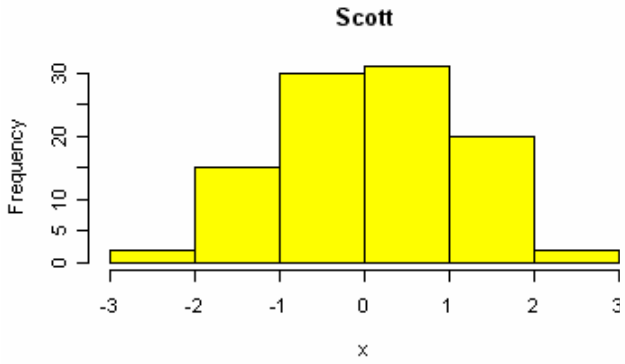
Histogrammes et nombre de classes

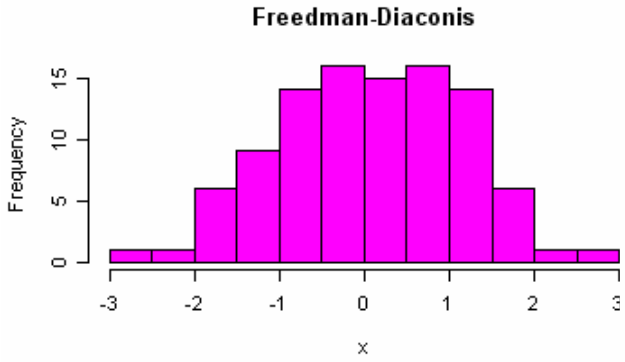
```
> x<-rnorm(100)
> NC <- function(x)c(Sturges=nclass.Sturges(x),Scott=nclass.scott(x),FD=nclass.FD(x))
> NC(x)
Sturges   Scott      FD
     15      49      64
> par(mfrow=c(3,1))
> hist(x,col="grey",main="Sturges")
> hist(x,col="yellow",breaks="Scott",main="Scott")
> hist(x,col="magenta",breaks="FD",main="Freedman-Diaconis")
```

Histogrammes et nombre de classes

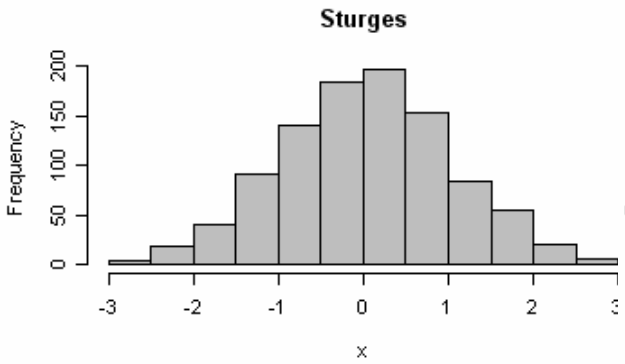
n = 100

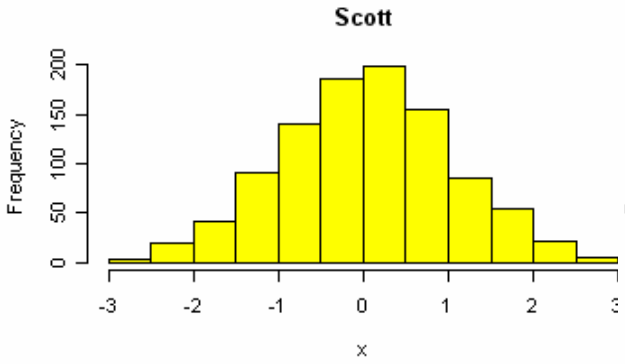


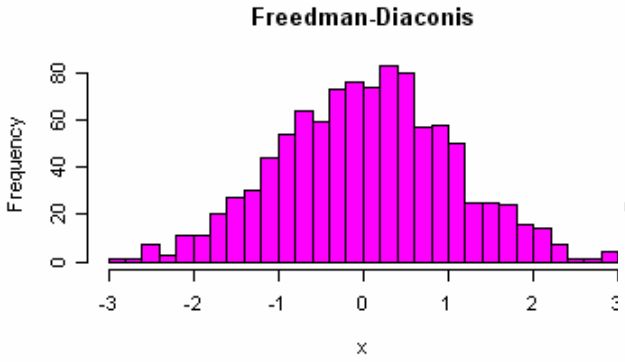




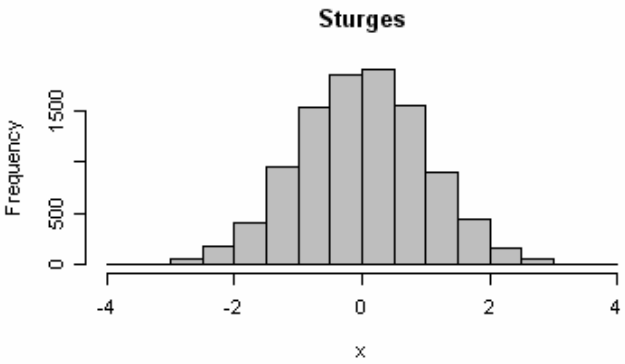
n = 1 000

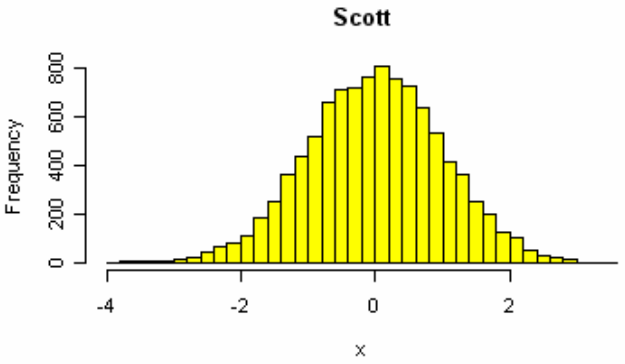


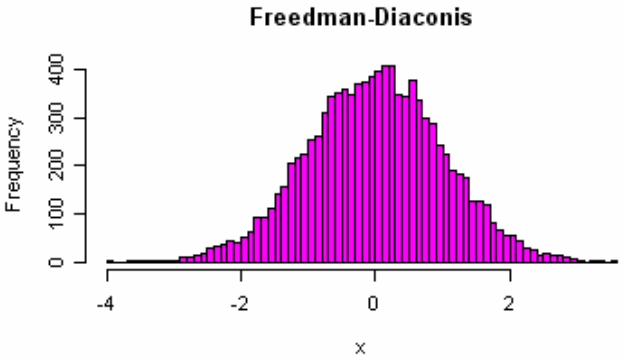




n = 10 000

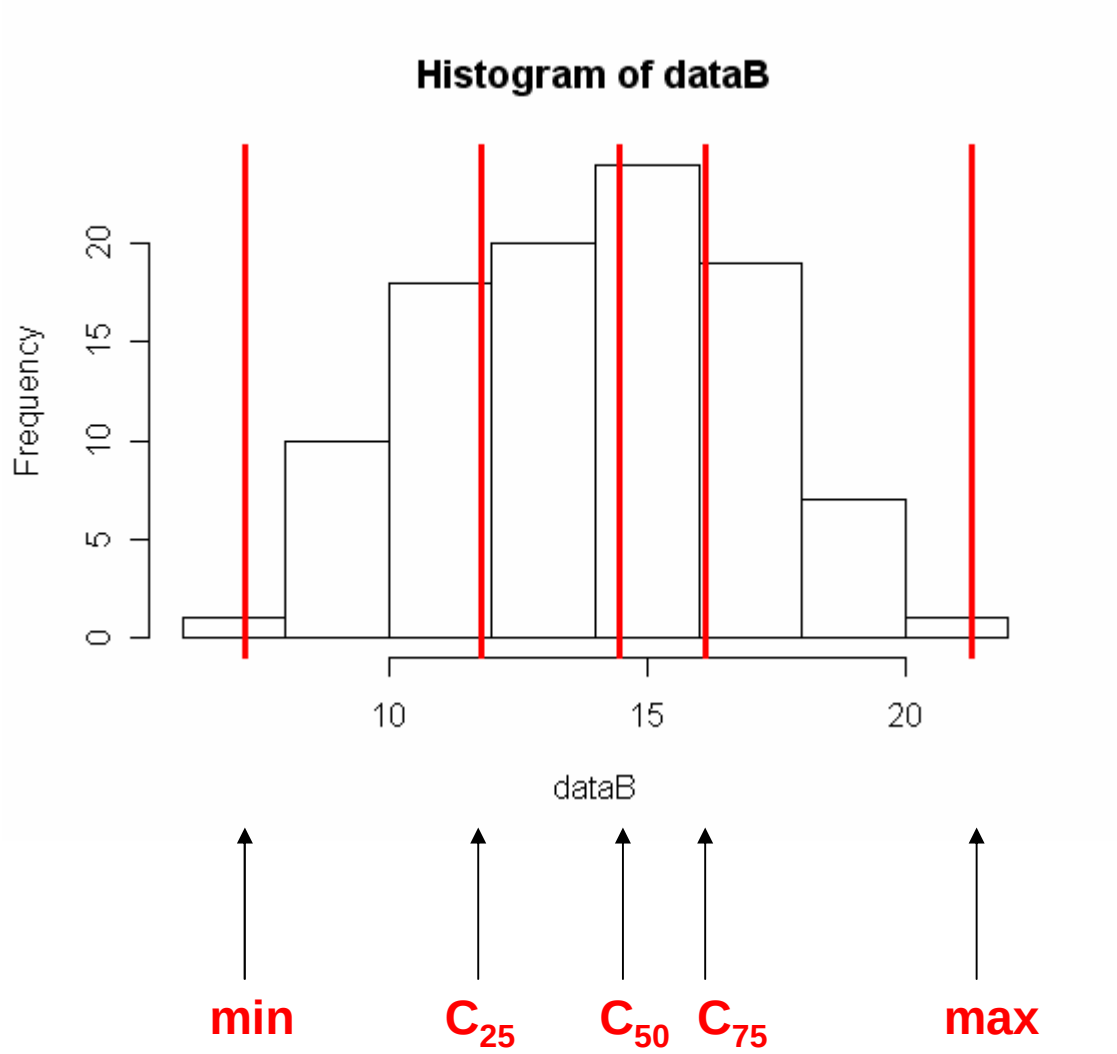






Histogrammes et mesures de position

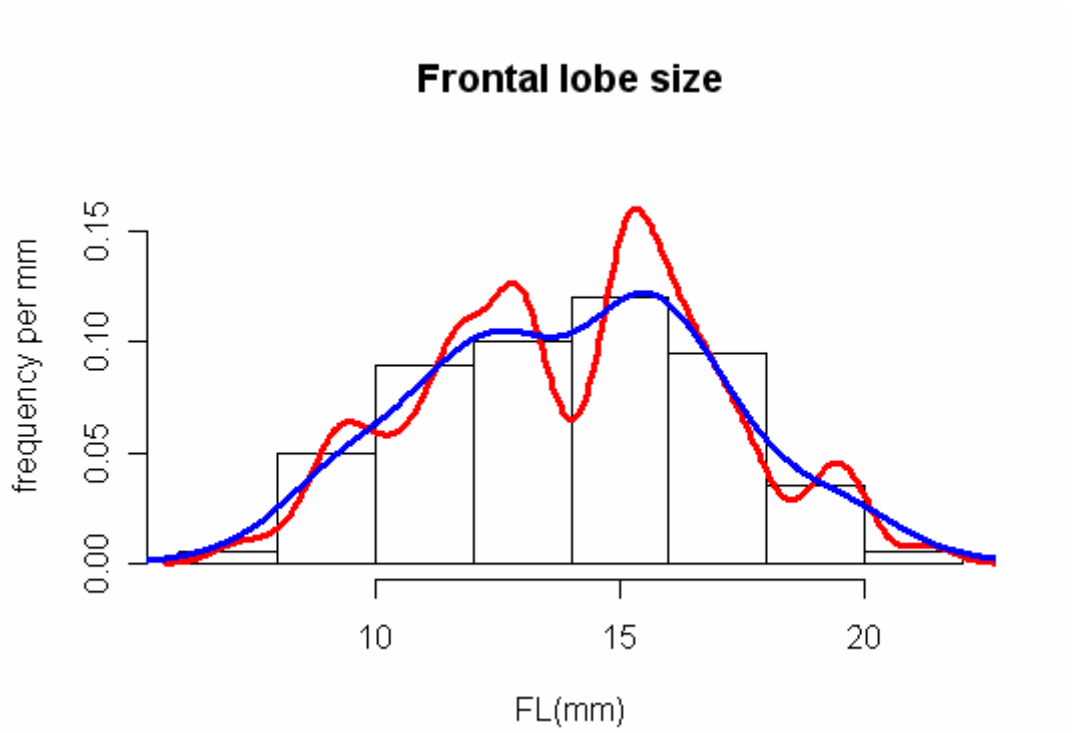
```
> hist(dataB)
> abline(v=quantile(dataB),col="red",lwd=3)
```



Histogrammes et courbe de densité

density

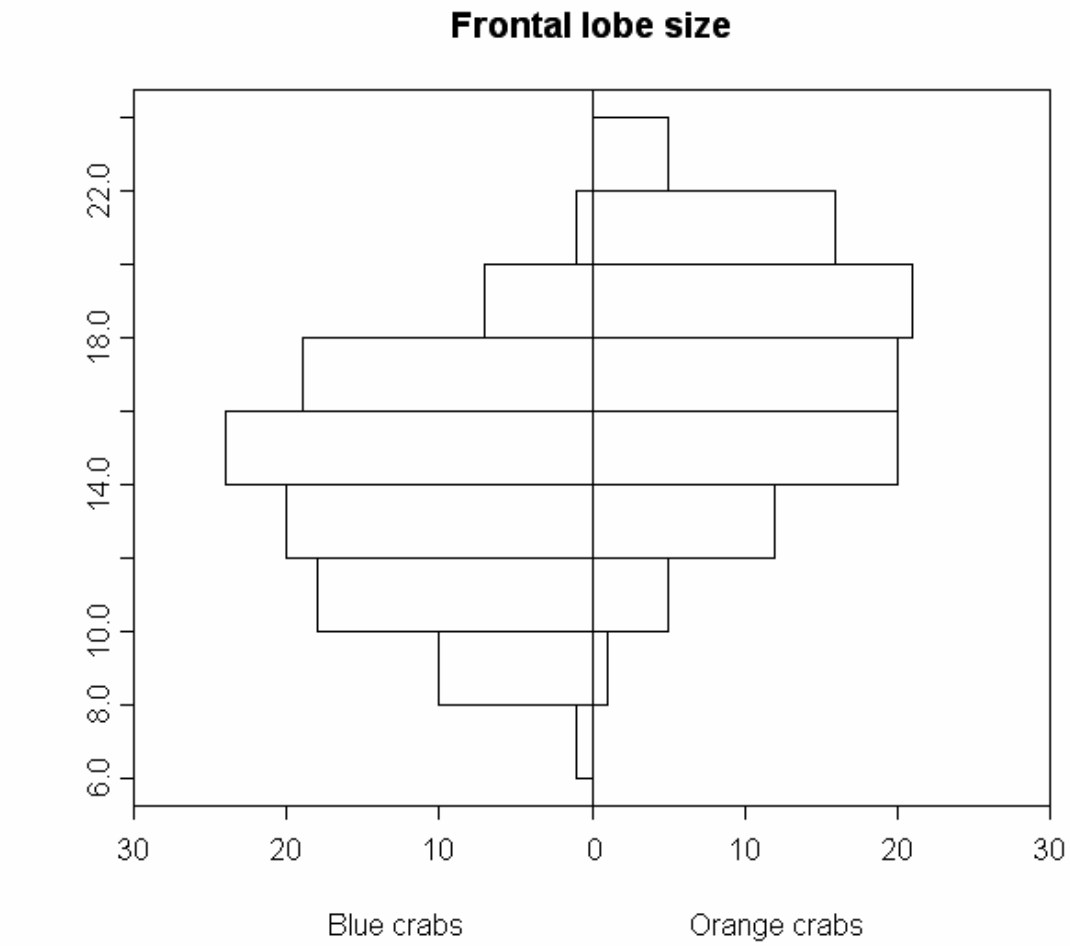
```
> hist(dataB,probability=TRUE,ylim=range(0,0.18),main="Frontal lobe size",
xlab="FL(mm)",ylab="frequency per mm")
> lines(density(dataB), col='blue',lwd=3)
> lines(density(dataB, bw=0.5), col='red',lwd=3)
```



Histogrammes multiples

histbackback

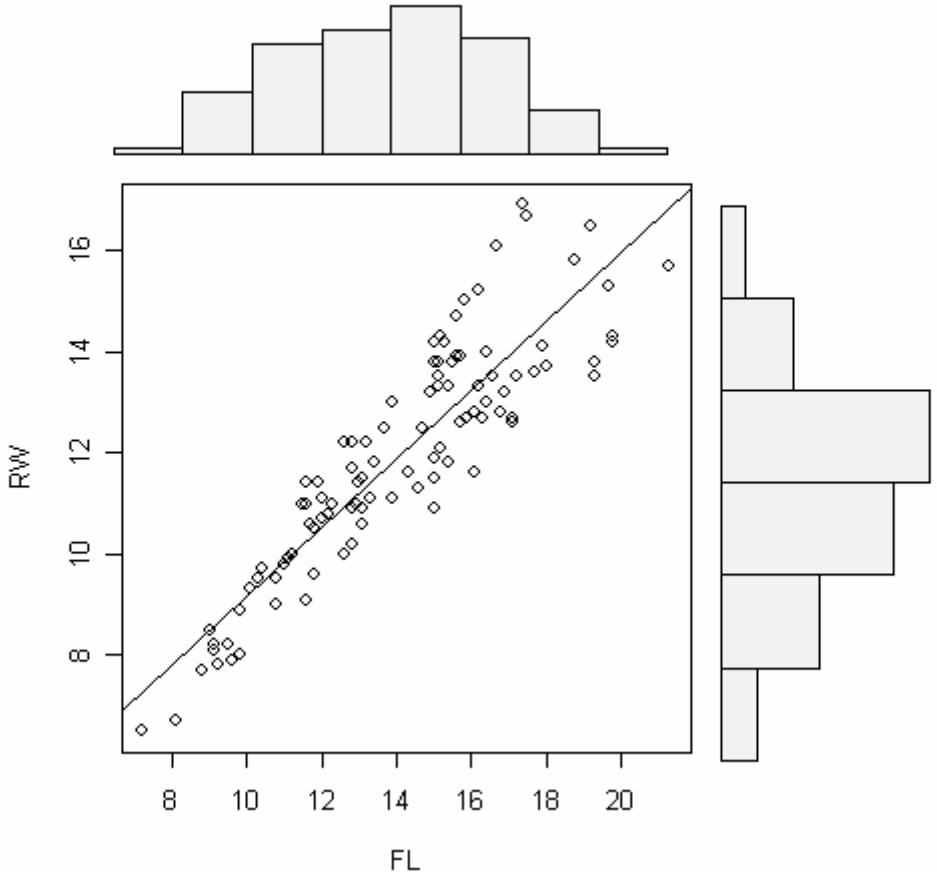
```
> library(Hmisc)
> options(digits=1)
> histbackback(dataB,dataO,xlab=c("Blue crabs","Orange crabs"),
main="Frontal lobe size")
```



Histogrammes et nuages de points

scatter.with.hist

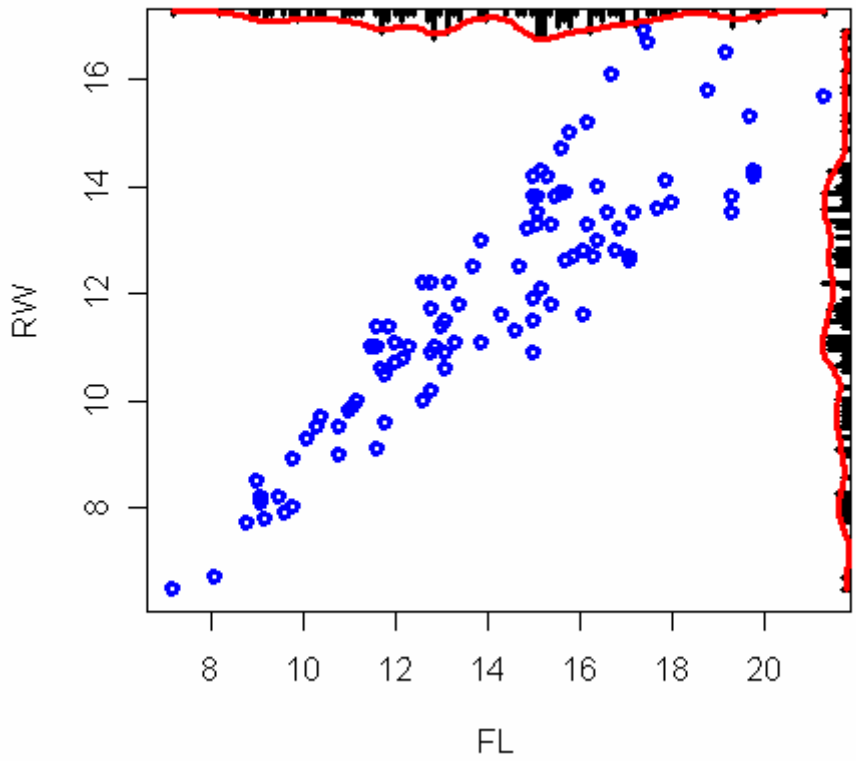
```
> library(UsingR)
> FL<-crabs[which(crabs$sp=="B"),4]
> RW<-crabs[which(crabs$sp=="B"),5]
> scatter.with.hist(FL,RW)
```



Histogrammes / courbes de densité et nuages de points

histSpike

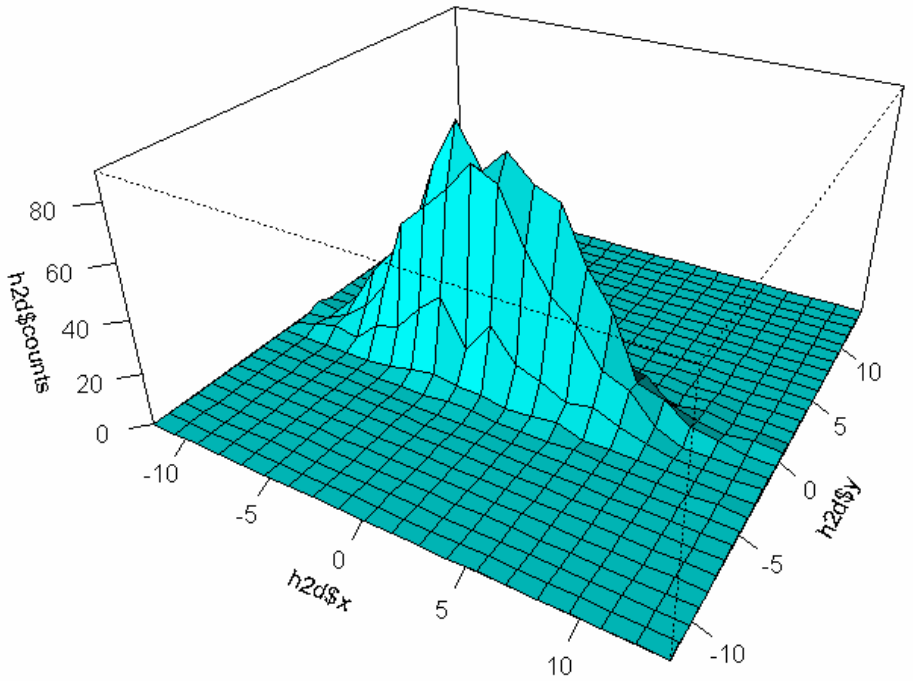
```
> library(Hmisc)
> plot(FL,RW,lwd=3,col="blue")
> histSpike(FL,3,add=TRUE,lwd=3)
> histSpike(RW,4,add=TRUE,lwd=3)
> histSpike(FL,3,type="density",col="red",lwd=3,add=TRUE)
> histSpike(RW,4,type="density",col="red",lwd=3,add=TRUE)
```



Histogrammes 2D

hist2d

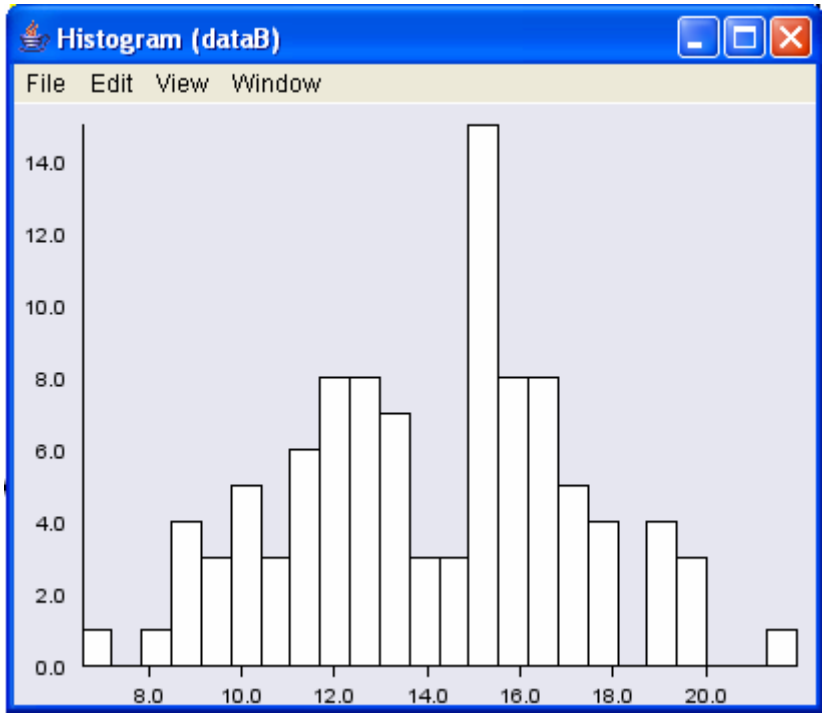
```
> library(gplots)
> x <- rnorm(2000, sd=4)
> y <- rnorm(2000, sd=1)
> h2d <- hist2d(x,y,show=FALSE, same.scale=TRUE, nbins=c(20,30))
> persp( h2d$x, h2d$y, h2d$counts,ticktype="detailed", theta=30, phi=30,
expand=0.5, shade=0.5, col="cyan",ltheta=-30)
```

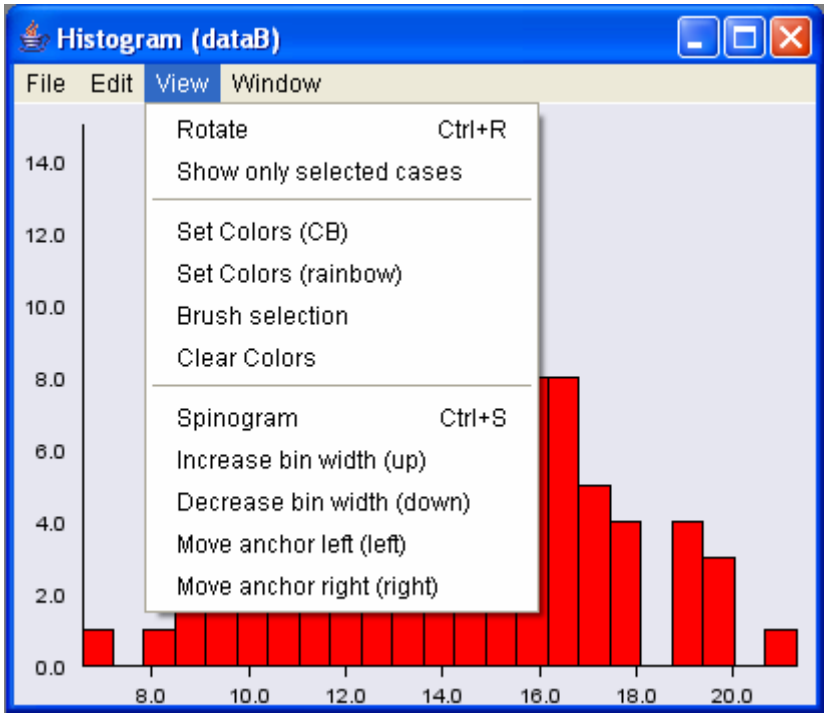


Histogramme interactif

ihist

```
> library(iplots)
> ihist(dataB)
```

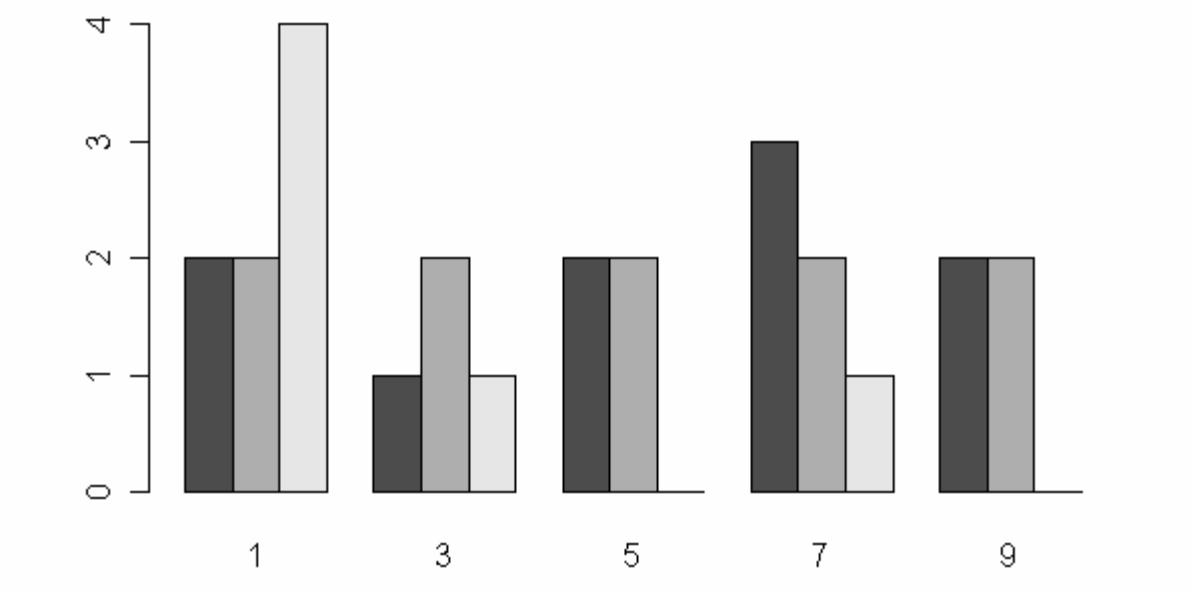




Histogrammes multiples

multhist

```
> library(plotrix)
> l <- list(runif(10)*10,1:10,c(1,1,1,1,4,8))
> l
[[1]]
 [1] 4.2947159 8.1528581 7.3732668 9.7694257 6.5140932 7.5774233 0.3487045
 [8] 4.3929983 2.2556319 1.8013953

[[2]]
 [1]  1  2  3  4  5  6  7  8  9 10

[[3]]
[1] 1 1 1 1 4 8
> multhist(l)
```



Estimation de la densité

Soient $x_1, x_2, \ldots, x_n$ n observations d'une variable aléatoire continue. On peut estimer la densité de cette distribution par la fonction :

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

h : largeur de fenêtre
K : densité de probabilité

Cette approche consiste à utiliser une fenêtre mobile que l'on déplace sur l'axe des valeurs de x et à compter le nombre d'occurrences appartenant à cette fenêtre.
Cette méthode donne une estimation peu régulière. Si l'on veut une fonction lisse, il est possible de moyenner une fonction de densité connue (le noyau) le long des données observées.