Analyse des données

Introduction à R

Printemps 2010

Qu'est-ce que R ?

- R : environnement de programmation statistique interfaçable avec C, Fortran ...
- Langage interprété et orienté objet semblable au langage statistique S (ou S+)
- Exécution et sémantique proche du langage Scheme, variante du Lisp
- Lecture, manipulation, stockage de données
- Nombreuses méthodes statistiques
- Outils graphiques variés avec sortie écran ou fichier
- Notion de modules (packages)
- Logiciel libre disponible sous Windows, Unix, Linux, Macintosh : diffusé sous licence GNU [a]

a. Projet créé en 1984 pour développer un système d'exploitation complet et libre

Historique

- Années 1970 : développement de S par John Chambers & co au Bell labs.
- R a été initialement écrit par Robert Gentleman et Ross Ihaka du département de statistique de l'université d'Auckland pour illustrer l'enseignement des statistiques
- Binaires mis à disposition en 1993 (Statlib)
- Source en free Software en 1995
- Élargissement du groupe en 1997
- Participation active de nombreux chercheurs du domaine : croissance exponentielle

Installation Windows

- Récupérer le fichier `R-2.8.1-win32.exe`
  - Aller sur un site CRAN, par exemple `http://cran.r-project.org/`
  - `Download and Install R` -> Windows
  - `base`
  - Download R 2.8.1 for Windows (34 megabytes)
- Création d'un environnement de travail (optionnel)
  - Créer un répertoire R dont l'adresse complère sera nommée <repertoire R> dans la suite.
  - Décompacter dans ce répertoire le fichier `R.zip`.
  - Positionner une variable d'environnement `R_USER` avec la valeur <repertoire R>/etc.
  - Dans le fichier <repertoire R>/etc/.Rprofile, mettre à jour l'adresse située dans la commande `setwd`

Démarrer et Terminer

- Démarrer :
  - Icône `R` (ou ligne de commande : `% R`)
  - R est exécuté dans le répertoire spécifié par la ligne de commande ou associé à l'icône de lancement
  - Lecture du fichier `.RProfile`
  - Lecture du fichier `.RData`
  - Execution de la fonction `.First`
  - Pour connaître le répertoire courant : `> getwd()`
- Quitter
  - `> q()`
  - Sauvegarde de l'environnement de travail dans le fichier `.RData`

L'aide de R

- Aide en ligne : `help.start()` ou par le menu : Aide -> Aide Html
- Aide pour une commande : `?commande` ou `help(commande)`
- Recherche par mots-clés : `apropos(mot-clé)`
- Commandes disponibles dans le menu `Aide`
- Documentation : officielle, contribuée, newsletters, livres, wiki, ...

Les objets

Eléments de base caractérisés par

- un nom
- une classe : vector, matrix, array, factor, time-series, data.frame, liste, fonction,...
- des attributs (par exemple le mode : logical, numeric, complex, character)

Création de vecteurs

- Concaténation

```
>  c(1,2,4)
[1] 1 2 4
> c("Alfred","George","Joseph")
[1] "Alfred" "George" "Joseph"
> c(T,F,T)
[1]  TRUE FALSE  TRUE
```

- Séquence : `seq(from,to)`, `seq(from,to,by=)`, `seq(from,to,length=)`

```
> seq(2,5)
[1] 2 3 4 5
> seq(2,5,by=0.5)
[1] 2.0 2.5 3.0 3.5 4.0 4.5 5.0
> seq(2,5,length=6)
[1] 2.0 2.6 3.2 3.8 4.4 5.0
```

- Répétition : `rep(x,times)`

```
> rep(c(2,3),3)
[1] 2 3 2 3 2 3
> rep(c(2,3),c(3,1))
[1] 2 2 2 3
```

Matrices et tableaux (matrice : tableau bidimensionnel)

- Un tableau est un vecteur muni d'un attribut dimension

```
> x<-seq(1,12)
> dim(x)<-c(3,4)
> x
     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
```

- Autre solution

```
> x<-matrix(1:12,nrow=3,byrow=T)
> x
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    5    6    7    8
[3,]    9   10   11   12
```

- Opérations usuelles : multiplication (%*%), transposition...

```
> t(x)
     [,1] [,2] [,3]
[1,]    1    5    9
[2,]    2    6   10
[3,]    3    7   11
[4,]    4    8   12
```

- Agrégation de matrices

```
> cbind(A=1:4,B=5:8,C=9:12)
     A B  C
[1,] 1 5  9
[2,] 2 6 10
[3,] 3 7 11
[4,] 4 8 12

> rbind(A=1:4,B=5:8,C=9:12)
  [,1] [,2] [,3] [,4]
A    1    2    3    4
B    5    6    7    8
C    9   10   11   12
```

Facteurs

Les facteurs sont des variables décrivant des catégories (souvent des variables qualitatives)

- Un facteur possède des niveaux (qui peuvent être ordonnés)

```
> pain <- c(0,3,2,2,1)
> fpain <- factor(pain,levels=0:3)
> levels(fpain) <- c("none","mild","medium","severe")
> fpain
[1] none   severe medium medium mild
Levels: none mild medium severe
```

Listes

Les liste sont des structures qui permettent de rassembler des données de types hétérogènes

- Création d'une liste

```
> maliste<-list(a=c(1,2),b="Alfred")
> maliste
$a
[1] 1 2

$b
[1] "Alfred"
```

Tableau de données : data.frame

Les tableaux de données rassemblent des vecteurs et facteurs décrivant des observations

- Création

```
> authors <- data.frame(
+ surname = c("Tukey", "Venables", "Tierney", "Ripley", "McNeil"),
+ nationality = c("US", "Australia", "US", "UK", "Australia"),
+ deceased = c("yes", rep("no", 4)))
> authors
   surname nationality deceased
1    Tukey          US      yes
2 Venables   Australia       no
3  Tierney          US       no
4   Ripley          UK       no
5   McNeil   Australia       no
```

Indexation

- Vecteurs

```
> x<-rbinom(10,prob=0.5,size=50)
> x
 [1] 16 27 31 24 22 25 23 25 27 30
> x[2]
[1] 27
> x[-2]
[1] 16 31 24 22 25 23 25 27 30
>  x[c(2,4)]
[1] 27 24
> x[x>25]
[1] 27 31 27 30
```

- Indexation d'une liste

```
>      maliste$a
[1] 1 2
>      listea<-maliste[1]
>      a<-maliste[[1]]
>      str(listea)
List of 1
 $ a: num [1:2] 1 2
>      str(a)
 num [1:2] 1 2
```

- Indexation de tableaux de données

```
> authors
   surname nationality deceased
1    Tukey          US      yes
2 Venables   Australia       no
3  Tierney          US       no
4   Ripley          UK       no
5   McNeil   Australia       no
> authors[2,]
   surname nationality deceased
2 Venables   Australia       no
> authors[authors$deceased=="yes",]
  surname nationality deceased
1   Tukey          US      yes
```

Valeurs manquantes

Les données comportent parfois des valeurs manquantes : expérience ratée, patient absent...

- `NA` est le code R pour les valeurs manquantes
- `is.na()` est une fonction booléenne qui permet de trouver les valeur manquantes

```
> x<-c(2,3,NA,3,5)
[1]  2  3 NA  3  5
> is.na(x)
[1] FALSE FALSE  TRUE FALSE FALSE
```

- De nombreuses fonctions traitent les valeurs manquantes spécialement :

```
> mean(x)
[1] NA
> mean(x,na.rm=T)
[1] 3.25
```

Ecrire une fonction

- Écrire la fonction directement dans R

```
> moyenne <- function(x,y)
+ {
+   m<-(x+y)/2
+   m
+ }
```

- Écrire la fonction dans un fichier texte `exemple.R` et la rendre connue du système

```
> source("exemple.R")
```

- Utiliser la fonction

```
> moyenne(3,6)
[1] 4.5
```

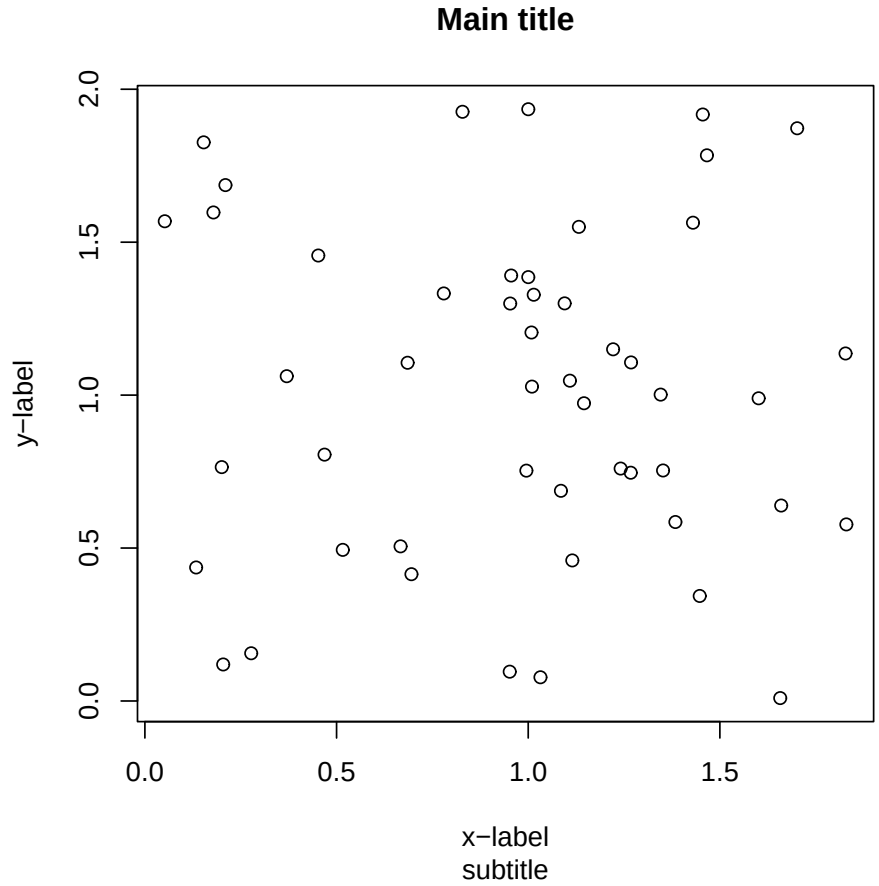
Exemple de fonction `.First`

```
.First<-function()
{
  print("Welcome to R")
  options(editor="emacs")
  }
```

Les fonctions graphiques

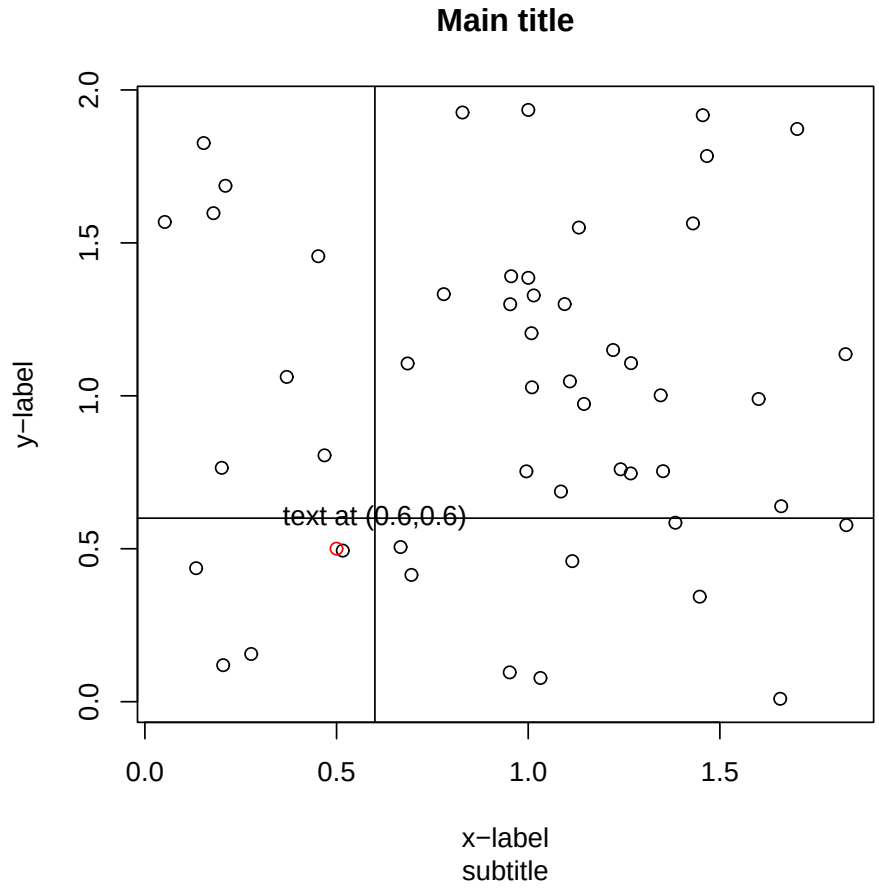
- Ouverture d'une figure

```
> x <- runif(50,0,2)
> y <- runif(50,0,2)
> plot(x, y, main="Main title", sub="subtitle",xlab="x-label", ylab="y-label")
```



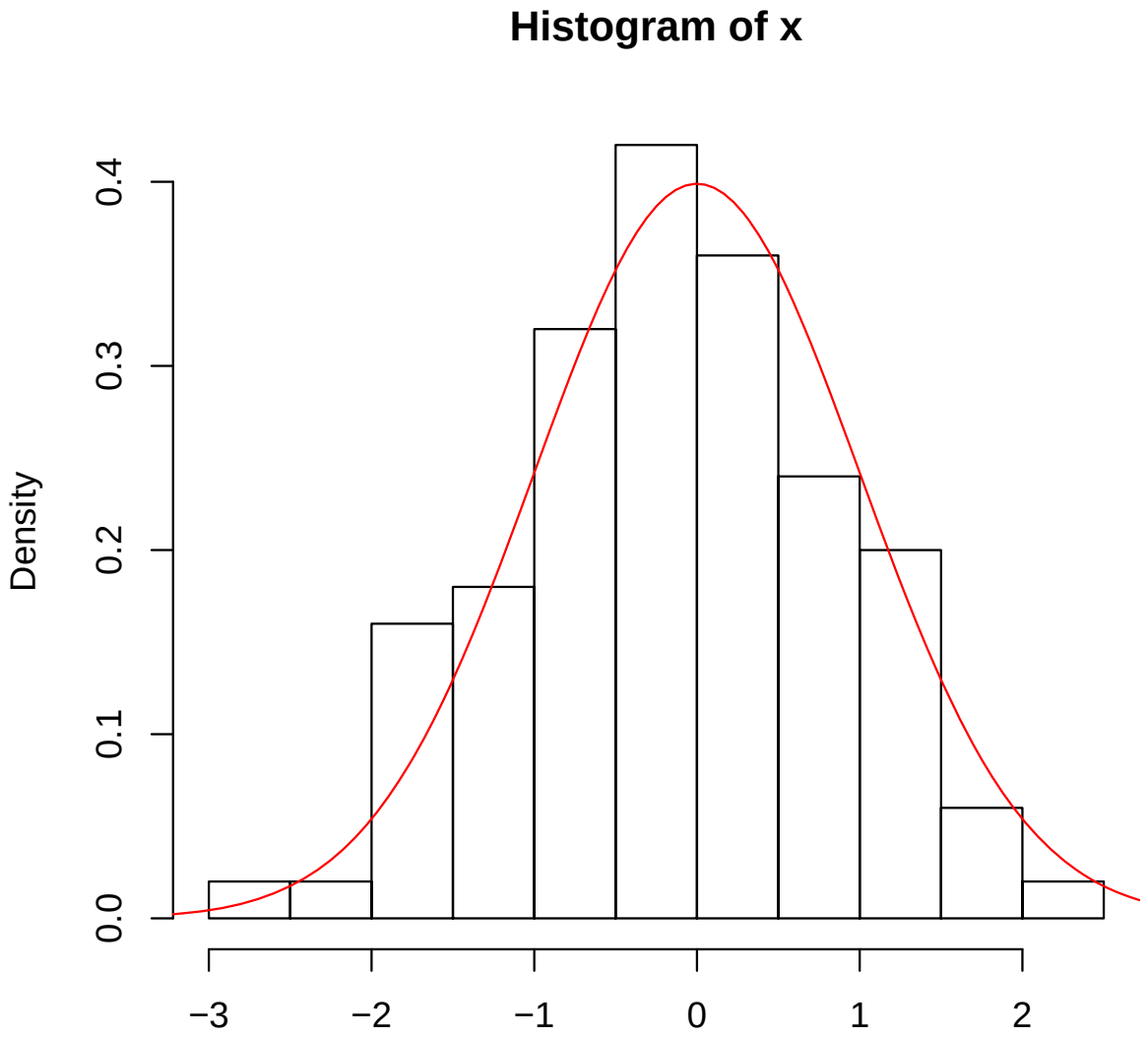
- Ajouts

```
> plot(x, y, main="Main title", sub="subtitle", xlab="x-label", ylab="y-label")
> text(0.6,0.6,"text at (0.6,0.6)")
> points(0.5,0.5,col="red")
> abline(h=.6,v=.6)
```



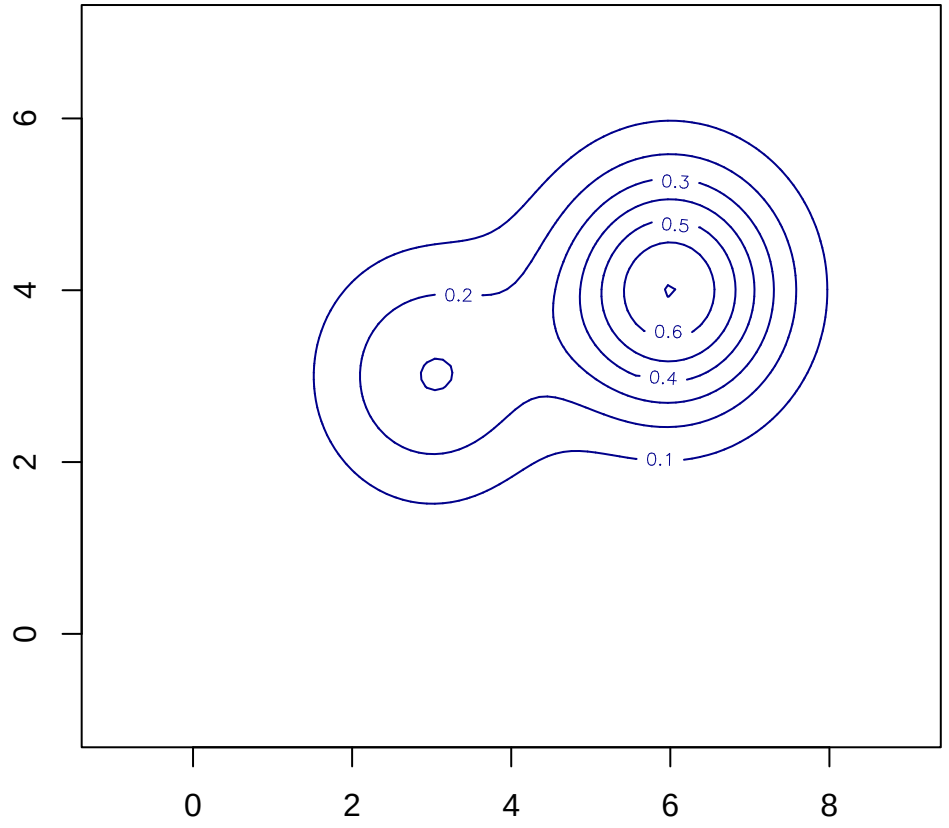
- Statistiques descriptives

```
> x <- rnorm(100)
> hist(x,probability=T)
> curve(dnorm(x),add=T,col=2)
```

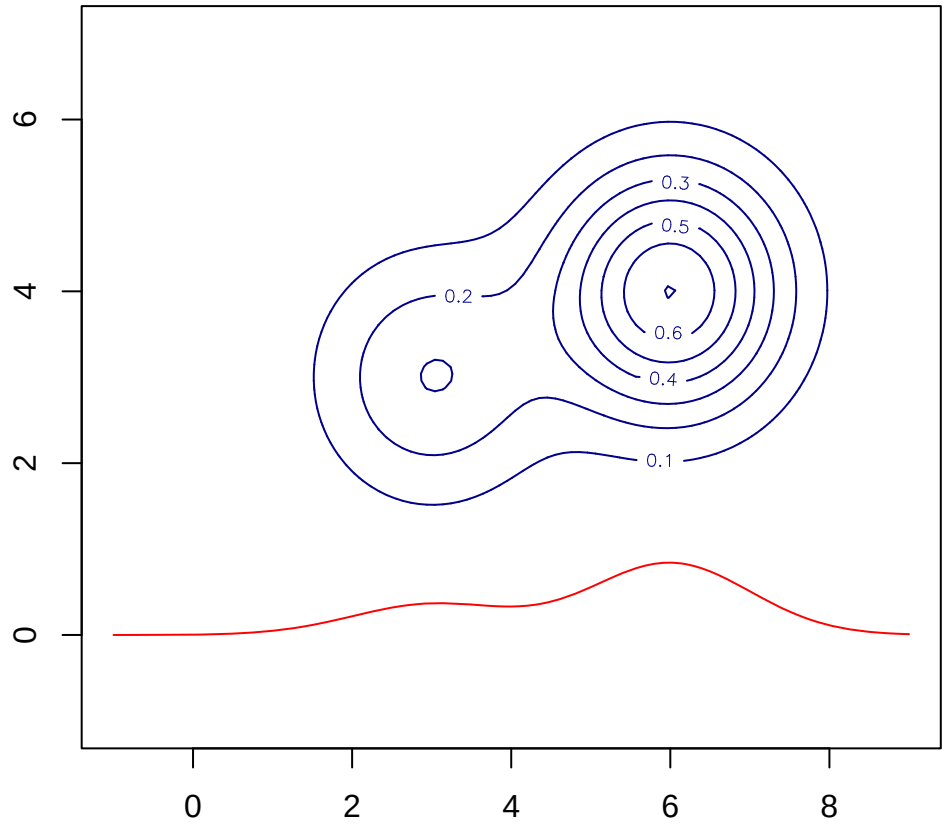


- Contours

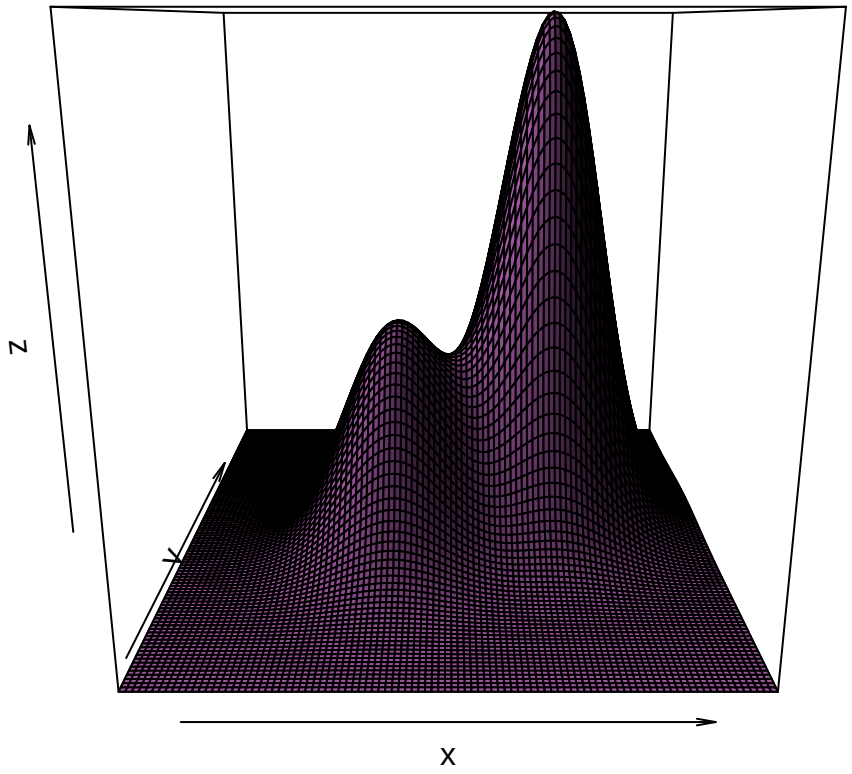
```
> x<-seq(-1,9,length=100)
> y<-seq(-1,7,length=100)
>  z<-outer(x,y,function(x,y) 0.3*exp(-0.5*((x-3)^2 +(y -3)^2)) +
> +   0.7*exp(-0.5*((x-6)^2 +(y -4)^2)))
> contour(x,y,z,col="blue4")
```



```
> contour(x,y,z,col="blue4")
> curve((0.3*dnorm(x,mean=3) + 0.7*dnorm(x,mean=6))*3,-1,9,col="red",ylim=c(-1,7),
    add=T)
```



```
> x<-seq(-1,9,length=100)
> lines((0.5*dnorm(x,mean=3) + 0.5*dnorm(x,mean=4))*sc,x,col="red")
> abline(h=0)
> abline(v=0)
> persp(x,y,z,shade=0.6,col="violet")
```



Les modules

- Modules : ensemble de fonctions
- Exemples de modules : `base`, `stats`,...
- `library()` donne la liste des modules installés
- `help(package=mod)` donne la liste des fonctions du module `mod`
- `library(mod)` pour charger un module installé
- `detach("package:mod")` pour décharger un module
- Commandes disponibles dans le menu `Packages`