Statistique Numérique et Analyse de Données
Ecole des Ponts ParisTech, 2ème année

TP1 : LOGICIEL R ET RAPPEL DES PROBABILITÉS

A propos du logiciel R

Le système R est un logiciel distribué gratuitement depuis le site

`http//www.r-project.org`

C'est un logiciel de type Open Source (Logiciel Libre en français) qui se développe grâce au bon-vouloir de certains contributeurs, soucieux de l'intérêt général. Du point de vue juridique, le logiciel est sous une licence GPL [1].

Le système R fournit un environnement intégrant un grand nombre de fonctionnalités statistiques et graphiques qui en font un outil particulièrement adapté au traitement et à l'analyse des données.
Un fichier executable permettant l'installation rapide de R sous Windows peut être téléchargé depuis le site `http//cran.us.r-project.org/bin/windows/base/`

- **Ouvrez une session sous Windows.**
- **Lancez le logiciel R.**
- **Essayez les commandes ci-dessous. Ne vous contentez pas à un simple copier-coller des commandes. Cherchez absolument à comprendre chacune des commandes utilisées. Si vous avez des doutes, n'hésitez pas à solliciter l'aide de l'enseignant.**

1. Les premiers pas sous R

1. Faire quelques essais :

```
pi*sqrt(10)+exp(4)
3:10
seq(3,10)
x = c(2,3,5,7,2,1)
y = c(10,15,12)
z = c(x,y)
z^2
```

1. General Public License - pour plus de détails sur les licences, consulter le site `http//www.gnu.org/licenses/license-list.fr.html`.

```
x*x
w=rep(x,3)
w=rep(x, each=3)
?rep     # Aide
ls()     # liste des variables saisies
rm(x)
x
ls()
```

2. Pour mieux connaître R :

```
?help
help(rep)
help(demo)
demo(graphics)
```

3. Passons aux matrices :

```
x = 1:12
dim(x) = c(3,4)
?dim
x
y = matrix(1:12, nrow=3, byrow=T)
t(y)
z = matrix(1:4, nrow=2, byrow=T)
z^2
z*z
z%*%z
```

4. Graphique :

```
x = runif(50, 0, 2)
y = runif(50, 0, 2)
plot(x, y, main="Titre", xlab="abscisse", ylab="ordonnée",col="darkred")
abline(h=.6,v=.6)
text(.6,.6, "placer un commentaire")
colors()
```

5. Définition de fonctions :
   - exemple très simple :

     ```
     carre = function(x) x^2
     carre(3)
     carre
     ```

   - exemple plus élaboré :

```
hist.norm=function(n, col)
{
x = rnorm(n)
h = hist(x, plot=F)
s = sd(x)
m = mean(x)
ylim = c(0,1.2*max(max(h$density),1/(s*sqrt(2*pi))))
xlab = "Histogramme et approximation par une loi normale"
ylab = " "
main = paste("Echantillon gaussien : n=",n)
hist(x, freq=F, ylim=ylim, xlab=xlab, ylab=ylab, col=col, main=main)
curve(dnorm(x,m,s), add=T, lwd=2)
}

op=par(mfcol=c(1,3))
hist.norm(200 , col="yellow")
hist.norm(800 , col="darkgoldenrod")
hist.norm(3200, col="blue")
par(op)
```

2. Simulation aléatoire

Dans les approches de modélisation, il est souvent utile de générer artificiellement des nombres (pseudo-)aléatoires.

1. Effectuer les essais suivants :

```
rnorm(10) # génère 10 réalisations de la loi N(0,1)
rnorm(10)
rnorm(10)
plot(rnorm(100))
rbinom(10, size=20, prob=.5)
rcauchy(10)
runif(10, min=0, max=1)
sample(1:40, 5)
sample(1:10, 10, replace=T)
sample(c("echec", "succes"), 10, replace=T, prob=c(0.7, 0.3))
```

2. Voici les loi les plus utilisées dont les tables statistiques sont intégrées dans `R` : `beta`, `binom`, `cauchy`, `chisq`, `exp`, `f`, `gamma`, `norm`, `pois`, `t`, `unif`.

3. Descriptions empiriques

1. Statistique d'ordre :

```
x = rnorm(10) # Echantillon i.i.d.
y = sort(x)   # Statistique d'ordre
```

2. Fonction de répartition empirique :

```
x = rnorm(100)
n=length(x)
plot(sort(x), 1:n/n, type="s", ylim=c(0,1), xlab="", ylab="")
?pnorm
curve(pnorm(x,0,1), add=T, col="blue")
```

3. Histogramme :

```
x = rnorm(100)
hist(x, breaks=20)
hist(x, breaks=20, freq=F, col="cyan")
curve(dnorm(x), add=T,col="darkblue")
x = rnorm(50)
h = hist(x, plot=F)
h$breaks
h$counts
?hist
```

4. Boxplot :

```
x = rnorm(100)
par(mfcol=c(2,2),bg="lightcyan")
boxplot(x)
boxplot(x,horizontal=T)
boxplot(x, col="red")
boxplot(x, col="orange",border="darkblue",lwd=2)
```

5. Boxplots en parallèle :

```
x = rnorm(100)
y = (rnorm(400))^2-1
z= rnorm(50)^3
par(bg="lightcyan")
boxplot(x,y,z,col=c("blue","white","red"),border=c("black","darkblue"),lwd=1.5)
```

6. QQ-plots :

```
x = rnorm(100)
y = (rnorm(400))^2-1
z = rnorm(200,m=4,sd=5)
par(bg="lightcyan",mfrow=c(2,2))
qqplot(x,y,pch=21,bg="red",fg="darkblue",lwd=2)
qqplot(x,z,pch=21,bg="red",fg="darkblue",lwd=2)
qqnorm(y,pch=21,bg="orange",fg="darkblue",lwd=2)
qqline(y,pch=21,col="blue",lwd=2)

qqnorm(z,pch=21,bg="orange",fg="darkblue",lwd=2)
qqline(z,pch=21,col="blue",lwd=2)
```

4. Lecture de données contenues dans un fichier

1. Les jeux de données sont généralement stockés dans des fichiers externes. La commande `read.table` permet de lire ce type de données. Pour tester cette commande,
   - téléchargez le fichier `AirQuality.data` de la page web du module
     `http://certis.enpc.fr/~dalalyan/StatNum.html`
     et placez-le dans votre répertoire de travail,
   - saisissez les commandes

```
Donnees = read.table("AirQuality.data")   # lire les données
summary(Donnees)                          # résumer les données
hist(Donnees$Ozone, col="gold") # histogramme de la variable Ozone
attach(Donnes)        # permet d ommettre le nom du jeu de données
hist(Ozone, freq=F, col="gold")
detach(Donnees)
hist(Ozone,freq=F,col="gold")   # erreur!
```

2. Un certain nombre de jeux de données sont fournis avec le logiciel R.

```
data()    # afficher la liste des jeux de données
?WWWusage # description des données WWWusage
plot(WWWusage)
```

5. Questions et exercices pour le compte-rendu

1. Dans la définition de la fonction `hist.norm`, pourquoi utilise-t-on la valeur `1/(s*¬ sqrt(2*pi))` ?

2. Produire des descriptions statistiques (moyenne, écart-type, médiane, min, max,...) des données réelles :

   ```
   library(MASS)
   data(geyser)
   attach(geyser)
   help(geyser)
   ```

   On pourra utiliser la commande `summary` ainsi que les commandes graphiques présentées précédemment (histogramme, fonction de répartition).

Le geyser Old Faithful

3. Commenter les boxplots en parallèle de la question 3.5. Laquelle des trois séries de données (a) est la plus dipersée ? (b) contient le plus grand nombre d'outliers (valeurs aberrantes) ?

4. Les données suivantes représentent les charges maximales (en tonnes) supportées par des câbles que fabrique une certaine usine :

10.1	12.2	9.3	12.4	13.7	10.8	11.6	10.1	11.2	11.3
12.2	12.6	11.5	9.2	14.2	11.1	13.3	11.8	7.1	10.5

   (a) Quelle est approximativement la valeur de la charge que les trois quarts des câbles peuvent supporter ?

   (b) Tracer le boxplot de ces données. Y a-t-il des valeurs aberrantes ? Dans ce diagramme, où visualise-t-on la valeur déterminée au point (a) ?

   (c) D'après le boxplot, la répartition de ces données semble-t-elle être symétrique ou pas ?

5. [2] Générer deux échantillons i.i.d. $U_1, \ldots, U_n$ et $V_1, \ldots, V_n$ selon la loi uniforme sur $[0, 1]$ (prendre n suffisament grand).

   i. A l'aide des représentations graphiques, deviner la loi des variables $T_i = -2 \log U_i$.

   ii. Etudier empiriquement la loi des variables $X_i = \sqrt{-2 \log U_i} \cos(2\pi V_i)$. D'après vous, quelle est cette loi ?

2. Les élèves en double diplôme n'ayant jamais suivi un cours de Théorie des Probabilités peuvent ne pas répondre a cette question.