HAI718 Probabilité et Statistiques

Introduction à R et lois usuelles

1 Commencer avec R

R est un système d'analyse statistique et graphique créé par Ross Ihaka et Robert Gentleman. Il dérive du dialecte S créé par AT&T Bell Laboratories. R est distribué librement, sous licence GNU GPL.

Pour lancer le logiciel R depuis les machines du RezUFR, il suffit de taper "R" dans un terminal. Pour quitter la session, il suffit de taper la commande `quit()`.

R est un langage interprété, et non compilé, ce qui signifie que les commandes tapées au clavier sont directement exécutées sans avoir besoin d'écrire un programme complet comme dans les langages comme C, Fortran ou Java.

La syntaxe de R est simple et intuitive, par exemple la fonction `rnorm(1)` génère une variable aléatoire de loi normale centrée réduite, `rnorm(1,2,5)` génère une variable aléatoire de loi normale $\mathcal{N}(2, 5)$. Les variables, données et fonctions du langages sont stockés dans la mémoire de l'ordinateur pendant tout le fonctionnement de la session, sous la forme d'*objets* portant chacun un *nom*. L'utilisateur agit sur ces objets avec des *opérateurs* (arithmétiques, logiques, de comparaison, ...) et des fonctions.

La fonction `ls()` permet de connaître, à un instant donné, les objets qui sont présents en mémoire.

1.1 Créér, lister et effacer des données en mémoire

L'assignation d'une valeur à une variable se réalise grâce à la flèche : `->`, dans un sens ou dans l'autre.

Exercice 1 *Observer ce que produisent les commandes suivantes tapées successivement :*

```
n <- 15
5 -> n
x <- 1
X <- 10
n
x
X
n <- 10 + 2
n
n <- 3 + rnorm(1)
(10 + 2)*5
name <- "Carmen"; n1 <- 10; n2 <- 100; m<- 0.5
ls()
```

La fonction `ls.str()` fournit des informations complémentaires sur les objets, et la fonction `ls(pat = "m")` va donner uniquement les objets dont les noms comportent un "m".

Exercice 2

1. *Testez les fonctions* `ls` *et* `ls.str` *avec différents patterns.*
2. *Que fait la sequence d'instructions suivante :*

```
M <- data.frame(n1, n2, m)
ls.str(pat = "M")
```

3. *Effacez tous les objets en mémoire à l'aide de la fonction* `rm`

1.2 L'aide en ligne

La fonction `help()` permet d'obtenir de l'aide sur la façon d'obtenir de l'aide... Pour résumer, il y a deux façons d'avoir une aide sur une fonction donnée, en tapant la commande précédée d'un point d'interrogation, ou en donnant la commande comme paramètre de l'aide. On peut aussi passer du mode aide en ligne au mode aide console en jouant sur le commutateur `htmlhelp={T,F}`.

Exercice 3 *Recherchez de l'aide sur la fonction* `rm`. *Comment détruit-on en une seule commande tous les objets de la session ? Comment détruit-on seulement les objets qui ont un "m" dans leur nom ? Testez l'aide en ligne* `help.start()`.

Dans ce qui suit, quand on spécifiera une fonction à utiliser pour résoudre un exercice, prenez l'habitude d'invoquer l'aide sur cette fonction pour comprendre son utilisation.

1.3 Tracer des graphiques

Lorsqu'une fonction graphique est utilisée en R, si aucune fenêtre graphique n'a déjà été créée, R en ouvre une automatiquement. Le dernier périphérique ouvert devient le périphérique actif dans lequel s'affichent les graphes suivants. La fonction `dev.list()` affiche la liste des périphériques ouverts. La liste des types de périphériques graphiques disponibles est accessible par `?device`.

Exercice 4 *Que font les commandes suivantes :*

```
x11(); x11(); pdf();
dev.list()
dev.cur()
dev.set(3)
dev.cur()
dev.off(2)
dev.list()
dev.off()
dev.off()
```

Les tracés de graphes se font à l'aide différentes fonctions. Nous utiliserons essentiellement `plot` et `hist`.

Exercice 5 *Après avoir lu l'aide relative à ces deux fonctions, définissez deux vecteurs de 1000 valeurs* `x <- rnorm(1000)` *et* `y <- rnorm(1000)`, *et tracez l'histogramme des fréquences de* `y` *puis les valeurs de* `3x+5` *en fonction des valeurs de* `x`. *Testez ensuite, en enlevant puis remettant chaque paramètre pour voir à quoi il correspond :*

```
plot(x,y,xlab="Mille valeurs au hasard", ylab="Mille autres valeurs",
     xlim=c(-2,2), ylim=c(-2,2), pch=22, col="red", bg="yellow", bty="l",
     tcl=0.4, main="Configurer les graphiques en R", las=1, cex=1.5)
```

Remarque : On peut avoir des explication sur les paramètres de type graphiques en invoquant l'aide `?par`.

Pour mettre deux graphiques sur la même courbe, on peut utiliser `points` et `lines`

Exercice 6 *Mettre sur un même graphique en bleu l'histogramme des propbabilités de* `x` *(paramètre* `probability=T` *de* `hist`*) et en rouge sa densité (fonction* `density`*).*

2 Lois binomiale, normale, de Student

2.1 La loi binomiale $\mathcal{B}(n, p)$

Soit $X \sim \mathcal{B}(n, p)$ avec $n = 18$ et $p = 1/6$.

Exercice 7 *Calculer $P(X = 3)$, $P(X \leq 3)$, $P(X \geq 3)$, $P(X \leq 16)$ à l'aide de la fonction pbinom().*

2.2 La loi normale $\mathcal{N}(m, \sigma)$

Exercice 8

1. *Soit $U \sim \mathcal{N}(0, 1)$ la loi normale centrée réduite.*
   (a) *À l'aide de la fonction pnorm(), calculer $P(U < 1.41)$, $P(U < -2.07)$, $P(U > -1.26)$.*
   (b) *À l'aide de la fonction qnorm(), trouver la valeur de u telle que $P(U < u) = 0.95$, $P(U < u) = 0.1$, $P(U > u) = 0.99$.*
2. *Soit $X \sim \mathcal{N}(m, \sigma^2)$ avec $m = -5$ et $\sigma = 4$.*
   (a) *Calculer $P(X < -5)$, $P(X \leq 0)$, $P(X \geq 5)$.*
   (b) *Trouver la valeur de x telle que $P(X < x) = 0.95$, $P(X < x) = 0.05$, $P(X > x) = 0.01$.*

2.3 La loi du Chi-deux χ^2 (ou loi de Pearson)

$U_1, \ldots, U_p$ étant p variables $\mathcal{N}(0, 1)$ indépendantes, on appelle loi du chi-deux à p degrés de liberté (χ^2_p) la loi de la variable $\sum_{i=1}^{p} U_i^2$.

Exercice 9

1. *Soit $X \sim \chi^2_{15}$ et $Y \sim \chi^2_{10}$. À l'aide de la fonction* `pchisq`*, calculer $P(X < 6.26)$, $P(Y > 3.25)$, $P(X + Y > 11.52)$.*
2. *Soit $X \sim \chi^2_{15}$. À l'aide de la fonction* `qchisq`*, trouver x tel que $P(X < x) = 0.01$, $P(X < x) = 0.05$, $P(X < x) = 0.99$.*

2.4 La loi de Student T_n

Soit une variable aléatoire $U \sim \mathcal{N}(0,1)$ et X une variable aléatoire suivant indépendamment de U une loi χ^2_n. On définit alors la variable de Student T_n à n degrés de liberté comme étant :

$$T_n = \frac{U}{\sqrt{\frac{X}{n}}}.$$

On note que la loi de Student T_n est symétrique, cela signifie que :

$$\forall x \ P(T_n < -t) = P(T_n > t).$$

De cette propriété, il découle que pour tout $x > 0$:

$$\text{si } P(|T_n| < t) = p \text{ alors } P(T_n < -t) = p/2 \text{ et } P(T_n > t) = p/2.$$

Exercice 10 *Soit $T \sim T_5$ une loi de Student à 5 degrés de liberté.*

1. *À l'aide de la fonction pt(), calculer $P(T < 0.408)$, $P(T < -2.07)$, $P(T > 0.132)$.*
2. *À l'aide de la fonction qt(), trouver la valeur de t telle que $P(T < t) = 0.05$, $P(T > t) = 0.9$, $P(T < t) = 0.5$.*

3 Simulation des lois binomiale, normale, de Student

3.1 La loi binomiale $\mathcal{B}(n,p)$

Exercice 11 *Soit $X \sim \mathcal{B}(n,p)$ avec $n = 20$ et $p = 0.04$.*
À l'aide de la fonction rbinom(), simuler un échantillon de taille 100 et 100000 de la loi de X et représenter les histogrammes pour chaque échantillon à l'aide de la fonction hist().

3.2 La loi normale $\mathcal{N}(m,\sigma)$

Exercice 12 *Soit $X \sim \mathcal{N}(m,\sigma)$ avec $m = -5$ et $\sigma = 4$.*

1. *À l'aide de la fonction rnorm(), simuler un échantillon de taille $n = 100$, et $n = 100000$ de la loi X et représenter les histogrammes pour chaque échantillon.*
2. *À l'aide des fonctions dnorm() et points(), représenter la fonction de densité de la variable X sur l'histogramme.*

3.3 Comparaison de la $\mathcal{N}(0,1)$ et de la T_n

À l'aide de graphiques, on souhaite comparer la loi normale (centrée réduite) et la loi de Student à n degrés de liberté pour différentes valeurs de n ($n = 5$, $n = 30$). Pour cela, on utilisera les deux fonctions *dt()* pour calculer la densité d'une loi de Student et *pt()* pour calculer une probabilité.

Exercice 13 *On comparera sur un même graphique les fonctions de densité pour la loi normale et les lois de Student. Faire de même avec les fonctions de répartition (la fonction de répartition de la variable X est la fonction $F(x) = P(X \leq x)$). Qu'en déduisez-vous ?*

4 Le théorème centrale limite

Théorème 1 *Soit $(X_i)_{i=1,...,n}$ une suite de variables aléatoires indépendantes, de même loi, de moyenne μ et de variance σ^2. On note $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$. On alors :*

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \text{ converge en loi vers } \mathcal{N}(0,1) \text{ quand } n \text{ tend vers l'infini}$$

On peut ainsi approximer la loi de $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ par la loi $\mathcal{N}(0,1)$ pour n assez grand ($n > 30$).

4.1 Application à la loi de Bernoulli

Soit $X_1, ..., X_n$ une suite de variables indépendantes de Bernoulli et de paramètre $p = 0.2$. On rappelle que l'espérance de X_i vaut p et que la variance est égale à $p(1-p)$.

Exercice 14 *On cherche à approximer la loi de $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ par une loi normale. Pour cela, on demande de simuler un échantillon de taille 1000 de la loi $\bar{X}$ pour $n = 10$, $n = 100$. Pour chacun des cas, on demande de tracer l'histogramme et la densité de la loi $\mathcal{N}(p, \frac{p(1-p)}{n})$.*