MASTER 1 SIDE **Pôle Lamartine - ULCO**

INFORMATIQUE ET MODÈLES LINÉAIRES Novembre 2011 - Session 1, Semestre 1

Durée de l'épreuve : 3h00 Tous documents autorisés.

(Les trois exercices sont indépendants. Un soin tout particulier sera apporté à la rédaction des réponses)

CORRECTION

Exercice 1

1. (a) En vous aidant de la commande `seq()`, générez la séquence

   -1, -0.9, -0.8,. . . , 0,. . . , 0.8, 0.9, 1

   et stockez la dans un vecteur x.

   ```
   > x<-seq (-1,1, by =0.1)
   ```

   (b) Comment extraire le sous-vecteur correspondant aux éléments situés aux positions 5 et 7 à 10 ?

   ```
   > x[c (5 ,7 :10)]
   ```

   (c) Comment extraire les éléments négatifs et les affecter au vecteur `xmoins` ?

   ```
   > xmoins<-x[x<0]
   ```

   (d) Comment retirer du vecteur `x` les éléments supérieurs à 0.5 et les affecter au vecteur `xnew` ?

   ```
   > xnew<-x[x<1/2]
   ```

2. (a) Générez une matrice M à l'aide de la fonction `matrix`, à 10 lignes et 5 colonnes, aléatoire (avec des valeurs réelles comprises entre 0 et 1).

   ```
   > M<-matrix(runif(50),nrow=10)
   ```

   (b) Déterminez le nombre d'éléments supérieurs à 0.9.

   ```
   > length(M[M >0.9])
   ```

   (c) Remplacez les éléments de M inférieurs à 0.5 par des 0.

   ```
   > M[M <0.5]<-0
   ```

   (d) Testez et vérifiez son type et la nature de ses éléments.

   ```
   > typeof(M)
   > class(M)
   ```

   (e) Créez un "data frame" nommé `MDF` à partir de M. Vérifiez que `MDF` a la structure voulue.

```
> MDF <-as.data.frame(M)
> typeof(M)
> class(M)
> class(MDF)
> typeof(MDF)
```

(f) Extrayez le vecteur correspondant à la troisième colonne.

```
> MDF[,3]
```

(g) Extrayez la liste correspondant à la deuxième ligne.

```
> MDF[2,]
```

3. Générez un data frame nommé `DF` tel que :
   (a) les variables (colonnes) sont nommées : "Sexe", "Âge" puis "Note 1", "Note 2",...,"Note 15" ;
   (b) le nombre d'individus (lignes) est de 50, et les lignes sont nommées "Étudiant n" où n est un numéro aléatoire unique (on utilisera les fonctions `row.names` et `paste`) ;
   (c) le sexe de chaque individu est choisi au hasard parmi "Masculin" et "Féminin" ;
   (d) les notes sont générées aléatoirement entre 0 et 20 et arrondies à 0.5 ;
   (e) Les âges sont générés aléatoirement entre 18 et 24.

```
> age<-sample(18 :24,40,replace=T)
> sexe<-sample(c("M","F"),40,replace=T)
> notes<-matrix(sample(0 :20,40*15,replace=T),nrow=40)
> DF<-data.frame(Age=age,Sexe=sexe,notes)
> row.names(DF)=paste(rep("Etud",40),as.character(sample(1 :40,40,replace=F)),
+ sep="")
```

Extrayez le sous-ensemble des données correspondant aux variables "Note 3", "Note 7", "Note 8",...,"Note 17". Extrayez le sous-ensemble des données correspondant aux filles.

```
> names(DF)[3 :17]<-paste(rep("Note",15),1 :15,sep="")
> DF.F<-DF[DF$Sexe=="F",] # premiere solution
> DF.F<-subset(DF,Sexe=="F") # deuxieme solution
```

Exercice 2 On s'intéresse au jeu de données stocké dans les fichiers "CO2.csv" ou "CO2.txt", présents dans le dossier MASTER1SIDE/Examen03-11-2011. Ce jeu présente le "Carbon Dioxide Uptake in Grass Plants" c'est-à-dire la consommation de dioxyde de carbone par des plantes herbeuses réfrigérées ou non ("Chilled" signifie "réfrigéré").

1. Ouvrez la table dans `R` à partir du fichier "CO2.csv" ou "CO2.txt" et visualisez la table, affichez le nom des variables colonnes. Présentez les différents types de plantes.

   (On n'oubliera pas de changer de répertoire courant...)

```
> CO2<-read.table(file="CO2.txt",header=TRUE)# ou
> CO2<read.csv(file="CO2.csv")
> CO2
> names(CO2)
> levels(CO2$Plant)
```

2. Précisez les modalités pour chacune des variables qualitatives (au nombre de 3) de la table.

```
> table(CO2$Plant)
> table(CO2$Type)
> table(CO2$Treatment)
```

3. Résumez l'information contenue dans la table.

```
> summary(CO2)
```

4. Représentez graphiquement à l'aide du graphique adéquat les variables qualitatives de la table dans une même fenêtre graphique.

```
> par(mfrow=c(1,3))
> pie(table(CO2$Plant))# ou
> barplot(table(CO2$Plant))
> pie(table(CO2$Type))# ou
> barplot(table(CO2$Type))
> pie(table(CO2$Treatment))# ou
> barplot(table(CO2$Treatment))
```

5. On s'intéresse à la colonne "conc". Retrouvez les informations contenues dans cette variable sans utiliser la fonction `summary`. Représentez graphiquement cette variable à l'aide du graphique adéquat.

```
> conc<-sort(CO2$conc)
> summary(conc)
> min(conc)
> max(conc)
> mean(conc)
> n<-length(conc)# n est pair ici donc on peut appliquer les formules ci-dessous
> Me<-(conc[n/2]+conc[n/2+1])/2
> Me
> Q1<-(conc[n/4]+conc[n/4+1])/2
> Q1
> Q3<-(conc[3*n/4]+conc[3*n/4+1])/2
> Q3
> hist(conc)
```

Exercice 3

On se donne l'énoncé ci-dessous issu d'un examen sur les modèles linéaires ainsi que sa correction. Retrouvez à l'aide de `R` les réponses aux questions.

On dit que dans une famille, les aînés ont tendance à être plus indépendants que leurs cadets. Un chercheur élabore une échelle d'indépendance en 25 points et procède à l'évaluation de 20 aînés et du frère ou de la sœur qui suit directement chacun des aînés. Imaginons qu'il obtienne les résultats suivants :

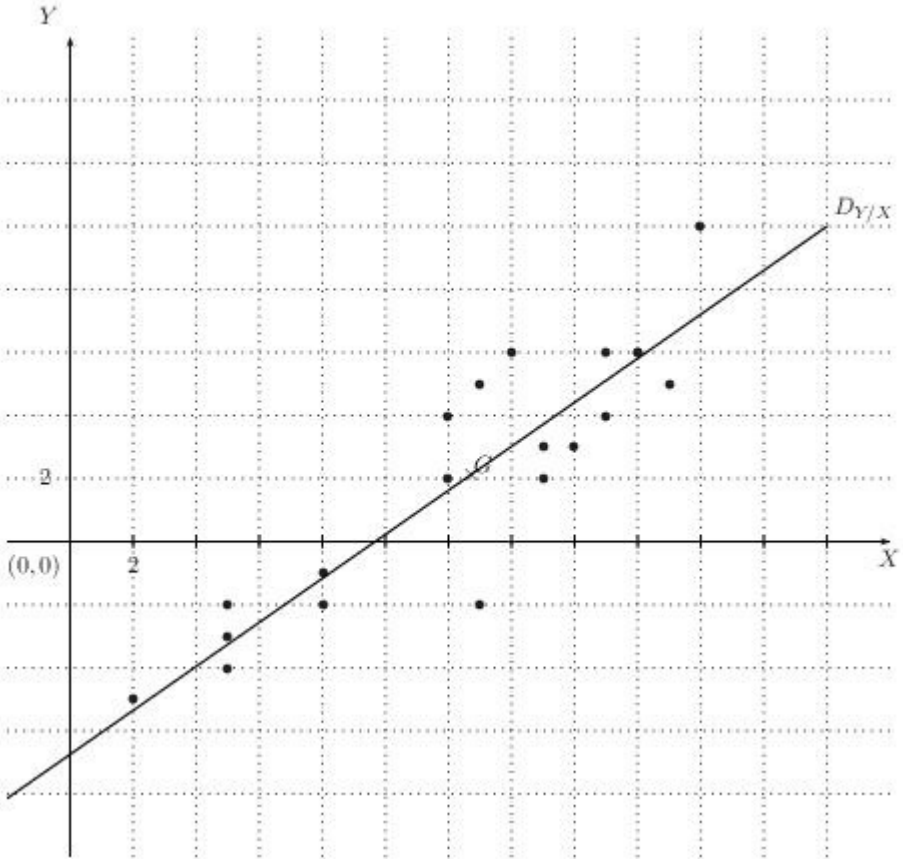
i	Aîné	Cadet	i	Aîné	Cadet
1	8	9	11	17	13
2	13	15	12	12	8
3	8	10	13	2	7
4	5	7	14	13	8
5	12	10	15	19	14
6	15	13	16	18	12
7	5	8	17	14	8
8	15	12	18	17	11
9	16	13	19	18	12
10	5	9	20	20	10

Un collaborateur du premier chercheur suggère que la différence observée sur une paire dépend essentiellement du score de l'aîné

1. Pour chaque paire, on appelle x_i le score de l'aîné et y_i la différence (algébrique) des scores entre l'aîné et le cadet. Calculez le coefficient de corrélation R entre les deux séries étudiées.
2. À l'aide d'un test de Student-Fisher au seuil de $\alpha = 5\%$, montrer que l'hypothèse H_0 : « $R = 0$ » n'est pas acceptable et que par conséquent, la corrélation est significative. On utilisera pour cela la statistique
$$T = \sqrt{n-2}\frac{R}{\sqrt{1-R^2}}$$
qui suit une loi de Student-Fisher à $\nu = n-2$ degrés de liberté, ainsi que la table de l'annexe.
3. Déterminez une équation de la droite de régression des (y_i) par rapport aux (x_i).
4. Représentez sur un graphique le nuage de points $(x_i; y_i)$ et la droite déterminée au 3.
5. Donnez une estimation du niveau d'indépendance d'un cadet dont le frère (aîné) a obtenu une note de 22

Correction :

1. On a $\overline{x} = \dfrac{252}{20} = 12,6$, $\overline{y} = \dfrac{43}{20} = 2,15$, $\sigma_X^2 = \dfrac{37222}{20} - (12,6)^2 = 27,34$ (donc $\sigma_X = 5,23$), $V(Y) = \sigma_Y^2 = \dfrac{415}{20} - (2,15)^2 = 16,13$ (donc $\sigma_Y = 4,015$) et $Cov(X,Y) = \dfrac{920}{20} - (12,6 \times 2,15) = 18,91$. Finalement, $R = \dfrac{18,91}{5,23 \times 4,015} = 0,9002$.
2. On détermine ensuite la valeur observée de la statistique T soit $T_{obs} = \sqrt{n-2}\dfrac{R}{\sqrt{1-R^2}} = 8,76$. Pour un nombre de degrés de liberté (d.d.l.) égal à 18, un seuil $\alpha = 5\%$, et un test bilatéral, on obtient grâce à la table de la loi de Student Fisher, $T_{lu} = 2,101$. On remarque que $T_{lu} < T_{obs}$ donc on rejette l'hypothèse H_0 selon laquelle R est négligeable. Le coefficient R est donc significatif (la corrélation est significative).
3. Montrons que la droite de régression admet pour équation
$$D_{Y/X} \ : \ Y = 0,69X - 6,56.$$
Les coefficients de l'équation $Y = bX + a$ de la droite des moindres carrés sont donnés par
$$\begin{cases} b &= \dfrac{Cov(X,Y)}{V(X)} = \dfrac{18,91}{27,34} = 0,69 \\ a &= \overline{y} - b\overline{x} = 2,15 - 0,69 \times 12,6 = -6,56 \end{cases}.$$
4. On a le nuage de point suivants.



On a utilisé le tableau suivant qui présente x_i et y_i pour tout $i \in \{1, \ldots, 20\}$:

i	Aîné x_i	y_i	i	Aîné x_i	y_i
1	8	-1	11	17	4
2	13	-2	12	12	4
3	8	-2	13	2	-5
4	5	-2	14	13	5
5	12	2	15	19	5
6	15	2	16	18	5
7	5	-3	17	14	6
8	15	3	18	17	6
9	16	3	19	18	6
10	5	-4	20	20	10

5. Si on pose $X = 22$ dans l'équation trouvée précédemment, on trouve $Y = 0,69 \times 22 - 6,56 = 8,62$. Une estimation du niveau d'indépendance d'un cadet dont le frère (aîné) a obtenu une note de 22 est de $22 - 8,62 = 13,38$.