MASTER 2 SIDE **Pôle Lamartine - ULCO**

Méthodes informatiques et Statistiques Janvier 2012 - Session 1, Semestre 2

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

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

Exercice 1 On se donne les données statistiques ci-dessous :

id	sexe	âge	taille	niveau	département	UFR	nbFS	score
1	F	22	1,7	L3	75	SJAP	1	3
2	H	20	1,76	L3	92	SEGMI	2	
3	H			L3	78	SEGMI	2	2
4	F	25	1,65	M2	75	SJAP	3	1
5	F	30	1,62	M2	92	STAPS	1	3
6	H	22	1,79	L3	75	SJAP	0	3
7	H	20	1,86	L3	92	SEGMI	2	
8	F	21	1,65	L3	78	STAPS	4	2
9	F	25	1,65	M2	75	STAPS	0	2
10	F	340	1,62	M2	92	STAPS	2	4

Pour chaque personne numérotée de 1 à 10, on précise
- son identifiant (numéro compris entre 1 et 10),
- son sexe (F ou H),
- son âge (en années),
- sa taille (en mètres),
- son niveau d'étude (Licence 1,2,3, Master 1,2),
- son département géographique d'origine
- l'UFR (Unité de Formation et de Recherche) à laquelle cette personne appartient (SJAP=Sciences Juridiques Administratives et Politiques, SEGMI=Sciences Économiques, Gestion, Mathématiques et Informatique, STAPS=Sciences et Techniques des Activités Physiques et Sportives)
- le nombre de ses frères et sœurs
- son score sur 4 à un test de connaissances.

1. Utilisez la commande `data.frame` afin de stocker les données du tableau, dans un "data frame" noté "DF".
   (On remarquera qu'il y a des données manquantes dans le tableau...)

```
> DF<-data.frame(id=1..10,
+ sexe=c("F","H","H","F","F","H","H","F","F","F"),
+ âge=c(22,20,NA,25,30,22,20,21,25,340),
+ taille=c(1.7,1.76,NA,1.65,1.62,1.79,1.86,1.65,1.65,1.62),
+ niveau=c("L3","L3","L3","M2","M2","L3","L3","L3","M2","M2"),
+ département=c(75,92,78,75,92,75,92,78,75,92),
+ UFR=c("SJAP","SEGMI","SEGMI","SJAP","STAPS","SJAP","SEGMI","STAPS","STAPS","STAPS"),
+ nbFS=c(1,2,2,3,1,0,2,4,0,2),
+ score=c(3,NA,2,1,3,3,NA,2,2,4))
```

2. Affichez la colonne "niveau" de "DF" ainsi que le type et les modalités de cette variable.

```
> DF$niveau
> class(DF$niveau)
```

```
> levels(DF$niveau)
```

3. Comment accède t-on à la 3e ligne de “DF” ?

```
> DF[3,]
```

4. Après des études et vérifications complémentaires, on a pu déterminer les valeurs manquantes du tableau et en corriger d'autres :
   – la 2e personne a obtenu un score de 4/4 à son contrôle de connaissances,
   – la 3e personne a 23 ans et mesure 1,70m,
   – la 7e personne a obtenu un score de 2/4 à son contrôle de connaissances,
   – la 10e personne n'a pas 340 ans mais 34 ans bien-sûr.
   Apportez les précisions et corrections précédentes à “DF”, tout en expliquant vos démarches.

```
> DF[2,9]<-4
> DF[3,3]<-23
> DF[3,4]<-1.7
> DF[7,9]<-2
> DF[10,3]<-34
```

5. Dans un certain nombre de cas, `R` n'a pas possibilité de donner le type correct d'une variable : il n'a aucun moyen d'identifier les variables ordonnées (il les prend pour des “factor”) car il ne connaît pas la relation d'ordre qui s'applique. C'est par exemple le cas de la variable “niveau”. De même, il ne peut pas identifier une variable nominale dont les modalités seraient des chiffres (comme les numéros de département). Nous allons donc devoir corriger ses choix. Pour transformer une variable numérique en facteur, il faut utiliser la fonction `as.factor`. Par exemple, `as.factor(DF$département)` permet de considérer la colonne “DF$département” non plus comme une variable numérique mais comme une nominale. La commande est la suivante :

```
> ### Modification du type de département
> DF$département<-as.factor(DF$département)
```

   Vérifiez que `R` décrit correctement le type de “DF$département” puis modifiez en conséquence les variables qui nécessitent également cette correction.

```
> class(DF$département)
> DF$département<-as.factor(DF$département)
> class(DF$département)
> class(DF$id)
> DF$id<-as.factor(DF$id)
> class(DF$id)
```

6. Pour chaque variable qualitative ou quantitative discrète, déterminez les effectifs associés.

```
> table(DF$sexe)
> table(DF$âge)
> table(DF$niveau)
> table(DF$département)
> table(DF$UFR)
> table(DF$nbFS)
> table(DF$score)
```

7. Pour les variables pour lesquelles il est possible de calculer les effectifs, tracez un diagramme en bâtons.

```
> plot(DF$sexe)
> barplot(table(DF$âge))
> barplot(table(DF$taille))
> plot(DF$niveau)
> plot(DF$département)
> plot(DF$UFR)
> barplot(table(DF$nbFS))
> barplot(table(DF$score))
```

8. Déterminez la médiane de la variable "taille". Comment interprétez-vous cette valeur ? Peut-on déterminer la médiane de la variable "UFR" ? Pourquoi ?

```
> median(DF$taille)
> # 50% des valeurs de la série relative à la variable ‘‘taille’’ sont inférieures (ou
supérieures) à 1,675.
> # On ne peut pas calculer la médiane de la variable ‘‘UFR’’ car cette variable n'est
pas quantitative.
```

9. Déterminez les quartiles de la variable "âge" de deux manières différentes. Interprétez ces 3 valeurs.

```
> # Première méthode
> quantile(DF$âge)
> # Seconde méthode
> âges.rangés<-sort(DF$âge)
> n<-length(âges.rangés)
> # n=10 est pair !
> Q2<-(âges.rangés[5]+âges.rangés[6])/2 # Deuxième quartile = médiane
> Q1<-âges.rangés[3]# Premier quartile
> Q3<-âges.rangés[8]# Troisième quartile
```

10. Estimez la dispersion des valeurs de la variable "score" autour de sa moyenne arithmétique.

```
> sd(DF$score)
> # La moyenne de la variable ‘‘score’’ est égale à 2.6. L'intervalle
> # [2.6-0.966 ;2.6+0.966]=[1.634 ;3.566] contient théoriquement 68%
> # des valeurs de la série.
```

11. Après avoir regroupé les valeurs de la variable "taille" dans des classes d'amplitude 0.1, présentez une représentation graphique des fréquences.

```
> hist(DF$taille,xaxp=c(1.6,1.9,3),breaks=3)
```

12. Représentez dans une seule fenêtre graphique les deux boîtes à moustaches de la variable "taille" des personnes ayant obtenu un score inférieur ou égal à 2 et de celles ayant obtenu un score supérieur ou égal à 3.

```
> par(mfrow=c(1,2))
> boxplot(DF$taille[DF$score<=2])
> boxplot(DF$taille[DF$score>=3])
```

13. Représentez dans une seule fenêtre graphique les deux diagrammes à secteurs circulaires de la variable "taille" des femmes et des hommes. Ajoutez-y des légendes et des titres.

```
> par(mfrow=c(1,2))
> pie(table(DF$taille[DF$sexe=="F"]))
> pie(table(DF$taille[DF$sexe=="H"]))
```

14. À partir de “DF”, est-il possible de crér deux nouveaux “data frame” spécifiques aux femmes (“DFF”) et aux hommes (“DFH”) ?

```
> # Sans aucun problème...
> DF.HF<-split(DF,DF$sexe)
> DF.HF$F
> DF.HF$H
```

Exercice 2

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.

```
> M[,3]
```

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

```
> M[2,]
```

Exercice 3

On s'intéresse au jeu de données stocké dans les fichiers "CO2.csv" ou "CO2.txt", présents dans le dossier MASTER SIDE/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)
```