ISFA 2ème année 2003-2004 - Analyse des données - Pratique

Contrôle sans machine. Merci d'utiliser l'espace imparti pour vos réponses.

1.

```
> w = 2
> cumsum(w)
```

Quelle est la réponse du logiciel **R** ? Votre choix :

a	b	c	d	e	f	g	h	i
[1] 2	[1] 1 2	2	1 2	4	[1] 4	[1] 1	Error	Autre

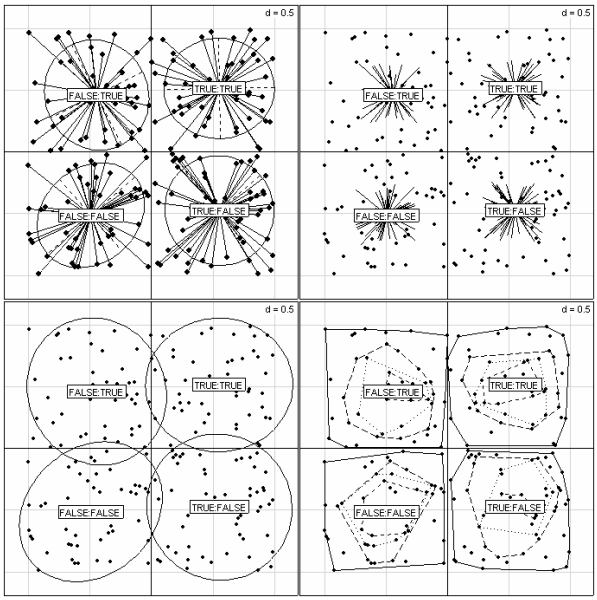
2.

```
> letters
 [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j" "k" "l" "m" "n" "o" "p" "q" "r" "s"
[20] "t" "u" "v" "w" "x" "y" "z"
> sum(sample(letters,260000,replace=TRUE)=="b")
```

Quel est l'ordre de grandeur du résultat ?

3.

Où trouve-t-on de l'information sur ce graphique ?



4.

```
> sum(sample(letters,26,replace=FALSE)=="b")==1
> all(rbinom(100,5,1/5))<6
> pcauchy(1.2999)<2
> pnorm(1)<0.975
> log(dpois(1,1))+1<1e-15
```

Lesquelles de ces assertions sont fausses ? (justifier).

5.

```
> a = matrix(1:8,2,4)
> b = matrix(8:1,2,4)
```

Indiquer dans la case correspondante ce que vous attendez des opérations `a+b`, `2*a` et `a*b` ?

6.

```
> a = matrix(1:8,2,4)
> b = matrix(8:1,2,4)
```

Indiquer dans la case correspondante ce que vous attendez des opérations `a==col(a)`, `round(a/b,dig=2)` et `a>b`.

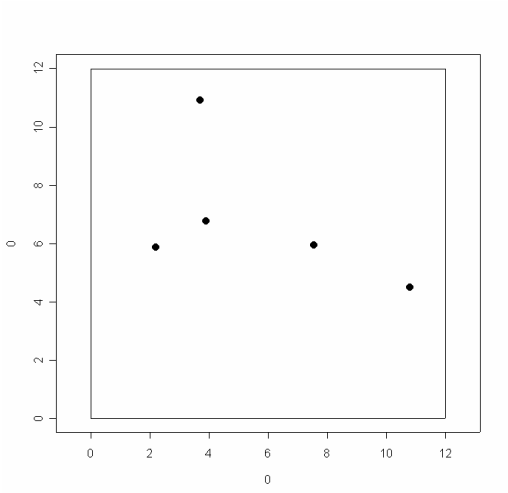
7.

```
> a = matrix(1:8,2,4)
> b = matrix(8:1,2,4)
```

indiquer dans la case correspondante ce que vous attendez des opérations `a%*%b`, `a$b` et `a%*%t(b)`

8.

Donner les instructions nécessaires pour dessiner un carré de sommets (0,0) et (12,12) avec 5 points choisis au hasard à l'intérieur, comme, par exemple :



9.

Dans la liste qui suit quel est l'intrus ?

```
is.data.frame(seconde)
is.numeric(seconde$HGEO)
is.factor(cut(seconde$HGEO,3))
is.list(seconde)
is.matrix(seconde[1:6,])
is.matrix(as.matrix(seconde))
is.character(names(seconde))
```

10.

```
> v1=c("toto","a",3.14159)
```

Indiquer dans la case correspondante ce que vous attendez des ordres suivants :

```
> v1
> v1[c(1,1,3,2,3,1)]
> v1(c(1,1,3,2,3,1))
> v1(c(1,1,3,2,3,1)]
> v1[1:4]
> v1[-2]
> v1[-(1:2)]
```

11.

Soit l'objet seconde de la librairie ade4. On y trouve la note moyenne en fin du premier trimestre de 22 élèves d'une classe de seconde pour 8 matières :

```
> data(seconde)
> seconde
   HGEO FRAN PHYS MATH BIOL ECON ANGL ESPA
1  11.6  8.7  4.5  7.6  9.0  6.5   10 15.5
2  13.6 12.3  6.2  8.5 12.0 11.5    7 12.0
```

```
3  13.2 12.1  8.5  6.3 11.6 11.0   13 14.5
4   8.8  8.2  5.0  3.7 10.6 10.0   11 14.5
5  12.7  8.6 10.0  9.3 10.6 14.0    9 13.5
6  12.4 10.0  5.5  9.2 10.5 11.0   10 15.0
7  12.6  9.3  9.5  9.6 10.5 12.5    9 11.0
8  14.4 14.0 15.0 18.8 11.4 15.0   14 14.5
9  14.0 10.6  7.2  8.8 12.2 15.0   12 19.5
10 13.2 13.1 18.0 12.3 12.4  9.5    9 13.0
11 10.6  8.0  4.5  7.1  9.8  8.0   14 16.5
12 13.4  8.8 13.0 13.8 12.6 10.0    9 11.5
13 14.0  9.0  8.7  7.7 11.1 13.5    9 12.6
14  9.7  9.6  4.0  5.5 12.0 12.0    7 14.0
15 15.2 13.8 11.0  8.7  8.0 15.5   14 12.8
16 13.4 11.8 11.0  9.5 13.6 13.5   11 12.9
17 11.4  8.8 11.0  7.2 12.2 12.5   11 13.0
18 11.2  8.8  3.7  8.1 11.8 10.5    5 14.0
19  9.0  8.2  5.7  8.3  8.5  5.0   10 14.0
20 11.4  9.5 10.0 10.2 12.8 13.5   10 13.0
21  9.7  9.6  3.7  9.1  9.5  6.0    8 13.5
22 13.6 10.2 12.0 15.8 13.2 10.0   13 14.5
```

Que faut-il faire pour obtenir :

1 - pour chacune des variables, la moyenne des 22 valeurs ;
2 - pour chacune des variables, la variance (en 1/*n*) des 22 valeurs ;
3 - pour chacune des variables, la variance (en 1/(*n* - 1)) des 22 valeurs ;
4 - pour chacune des variables, le minimum des 22 valeurs ;
5 - la matrice de corrélation entre les 8 variables.

12.

Dans la documentation de dudi.pca :

```
Description:

     'dudi.pca' performs a principal component analysis of a data frame
     and returns the results as objects of class 'pca' and 'dudi'.

Usage:

     dudi.pca(df, row.w = rep(1, nrow(df))/nrow(df),
         col.w = rep(1, ncol(df)), center = TRUE, scale = TRUE,
         scannf = TRUE, nf = 2)
```

```
> pca1=dudi.pca(seconde,scannf=FALSE)
> names(pca1)
 [1] "tab"  "cw"   "lw"   "eig"  "rank" "nf"   "c1"   "l1"   "co"   "li"
[11] "call" "cent" "norm"
```

La fonction renvoie une liste. Indiquer en quelques mots, dans l'espace prévu, la nature des composantes de cette liste ayant les noms tab, cw, lw, eig, rank et nf.

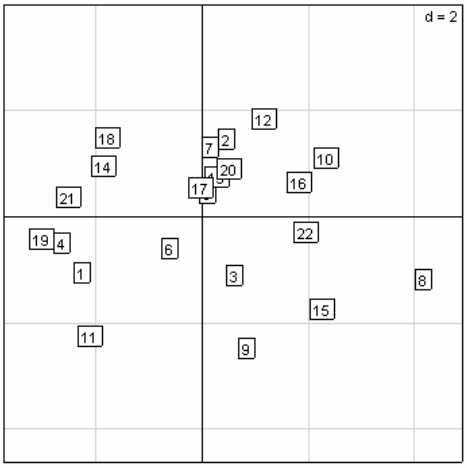
13.

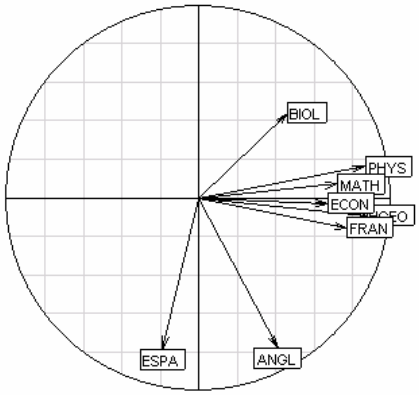
Indiquer en quelques mots dans l'espace prévu la nature des autres composantes de cette liste.

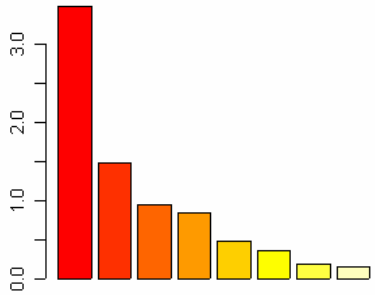
14.

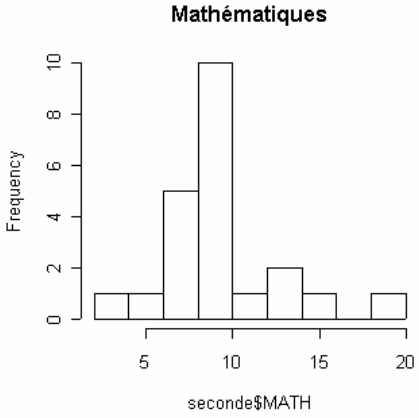
Quelles fonctions sont elles utilisées pour faire le graphe ci-dessous ? Indiquer si possible l'instruction utilisée. On a fait au préalable :

```
> par(mfrow=c(2,2))
```









15.

Donner le numéro du meilleur élève de la classe. Donner les numéros du groupe d'élèves ayant les moins bons résultats. Dites comment vous avez fait votre choix.

16.

Donner le numéro d'un élève assez bon dans les matières principales mais ayant des difficultés en langue. Donner le numéro d'un élève plutôt médiocre dans les matières principales mais ayant de bons résultats en langue.

17.

```
> pca1$eig
[1] 3.4926 1.4911 0.9596 0.8503 0.4834 0.3685 0.2015 0.153
```

Indiquer le résultat attendu des ordres suivants et justifiez :

```
1 -             > sum(pca1$eig)
2 -             > t(pca1$c1)%*%as.matrix(pca1$c1)
3 -             > t(pca1$li)%*%as.matrix(pca1$li)/22
4 -             > (pca1$co[1,1]/sqrt(pca1$eig[1])) == pca1$c1[1,1]
```

18.

Que valent les quantités suivantes ?

```
eigen(cor(pca1$tab))$values[2]
eigen(pca1$tab)$values[2]
```

19.

Que vaut la quantité suivante ?

```
svd(cor(pca1$tab))$d[2]
```

20.

Que vaut la quantité suivante ?

```
svd(pca1$tab/sqrt(22))$d[2]^2
```

ISFA 2ème année 2003-2004 - Solution

1.

```
> w = 2
> cumsum(w)
[1] 2
```

la réponse est a

2.

1 chance sur 26 et 260000 essais donnent un résultat autour de 10000. Par exemple :

```
sum(sample(letters,260000,replace=TRUE)=="b")
[1] 9782
```

3.

C'est le résultat d'un exemple de la fonction s.class de la librairie ade4.

4.

Elles sont toutes vraies.
sum(sample(letters,26,replace=FALSE)=="b")==1, c'est vrai.
Chaque lettre sort une fois et une seule, obligatoirement.
all(rbinom(100,5,1/5))<6, c'est vrai.
Une variable aléatoire binomiale (n=5, p = 1/5) prend ses valeurs dans {0, 1, 2, 3, 4, 5}.
pcauchy(1.2999)<2, c'est vrai.
Une fonction de répartition est toujours inférieure ou égale à 1
pnorm(1)<0.975, c'est vrai.
Le quantile 0.975 d'une loi normale est aux environs de 1.96
log(dpois(1,1))+1<1e-15, c'est vrai.
La valeur exacte de la probabilité de 1 pour une loi de Poisson de paramètre 1 vaut exactement :

$$P(X=1) = e^{-\lambda}\frac{\lambda^1}{1!} = \frac{1}{e}$$

Le log vaut donc -1 et la somme est nulle.

5.

```
> a+b
     [,1] [,2] [,3] [,4]
[1,]    9    9    9    9
[2,]    9    9    9    9
> 2*a
     [,1] [,2] [,3] [,4]
[1,]    2    6   10   14
[2,]    4    8   12   16
> a*b
     [,1] [,2] [,3] [,4]
[1,]    8   18   20   14
[2,]   14   20   18    8
```

6.

```
> a==col(a)
      [,1]  [,2]  [,3]  [,4]
[1,]  TRUE FALSE FALSE FALSE
[2,] FALSE FALSE FALSE FALSE
> round(a/b,dig=2)
     [,1] [,2] [,3] [,4]
[1,] 0.12  0.5 1.25  3.5
[2,] 0.29  0.8 2.00  8.0
> a>b
      [,1]  [,2] [,3] [,4]
[1,] FALSE FALSE TRUE TRUE
[2,] FALSE FALSE TRUE TRUE
```

7.

```
> a%*%b
Error in a %*% b : non-conformable arguments
Le produit de matrice est impossible vues les dimensions
>  a$b
NULL
On cherche la composante b de la liste a, qui n'existe pas
> a%*%t(b)
     [,1] [,2]
[1,]   60   44
[2,]   80   60
Le produit est possible, t est la transposition
```

8.

```
> plot(0,0,,type="n",xlim=c(0,12),ylim=c(0,12),asp=1)
> rect(0,0,12,12)
> points(runif(5,0,12),runif(5,0,12),cex=2,pch=20)
```

9.

```
Dans la liste l'intrus est
                    is.matrix(seconde[1:6,])
c'est le seul qui soit faux, car c'est un data.frame. Tous les autres sont vrais.
```

10.

```
> v1
[1] "toto"    "a"       "3.14159"
> v1[c(1,1,3,2,3,1)]
[1] "toto"    "toto"    "3.14159" "a"       "3.14159" "toto"
> v1(c(1,1,3,2,3,1))
Error: couldn't find function "v1"
> v1(c(1,1,3,2,3,1)]
Error: syntax error
> v1[1:4]
[1] "toto"    "a"       "3.14159" NA
> v1[-2]
[1] "toto"    "3.14159"
> v1[-(1:2)]
[1] "3.14159"
```

11.

```
> apply(seconde,2,mean)
  HGEO   FRAN   PHYS   MATH   BIOL   ECON   ANGL   ESPA
12.232 10.136  8.532  9.323 11.177 11.182 10.227 13.877

> apply(seconde,2,var)
  HGEO   FRAN   PHYS   MATH   BIOL   ECON   ANGL   ESPA
 3.288  3.500 15.413 11.234  2.309  8.608  5.898  3.228

> apply(seconde,2,var)*21/22
  HGEO   FRAN   PHYS   MATH   BIOL   ECON   ANGL   ESPA
 3.139  3.340 14.712 10.724  2.204  8.217  5.630  3.081

> apply(seconde,2,min)
HGEO FRAN PHYS MATH BIOL ECON ANGL ESPA
 8.8  8.0  3.7  3.7  8.0  5.0  5.0 11.0

> cor(seconde)
        HGEO     FRAN    PHYS    MATH    BIOL     ECON    ANGL     ESPA
HGEO  1.0000  0.67726  0.6317  0.5325  0.2704  0.65316  0.3292 -0.10165
FRAN  0.6773  1.00000  0.5567  0.4387  0.1578  0.45813  0.3324 -0.08745
PHYS  0.6317  0.55667  1.0000  0.6982  0.4151  0.41743  0.3323 -0.33990
MATH  0.5325  0.43868  0.6982  1.0000  0.3150  0.17921  0.2497 -0.13166
BIOL  0.2704  0.15779  0.4151  0.3150  1.0000  0.36310 -0.1431 -0.09371
ECON  0.6532  0.45813  0.4174  0.1792  0.3631  1.00000  0.2178 -0.06242
ANGL  0.3292  0.33244  0.3323  0.2497 -0.1431  0.21780  1.0000  0.38869
ESPA -0.1017 -0.08745 -0.3399 -0.1317 -0.0937 -0.06242  0.3887  1.00000
```

12.

tab	Le tableau centré réduit (22 lignes et 9 colonnes)
cw	le poids des colonnes (9 fois l'unité)
lw	le poids des lignes (22 fois 1/22)
eig	les valeurs propres de la matrice de corrélation
rank	le rang de la matrice de corrélation
nf	le nombre de facteurs conservés (par défaut 2)

13.

c1	les axes principaux (9 lignes et 2 colonnes)
l1	les composantes principales (22 lignes et 2 colonnes)
co	les coordonnées des colonnes (c1 * la valeur singulière)
li	les coordonnées des lignes (l1 * la valeur singulière)
call	l'ordre d'appel de la fonction qui a créé l'objet
cent	les moyennes
norm	les écrts-types en 1/n

14.

```
> par(mfrow=c(2,2))
> s.label(pca1$li)
> s.corcircle (pca1$co)
> barplot(pca1$eig)
> hist(seconde$MATH,main="Mathématiques")
```

15.

Sur la carte factorielle des lignes, en haut à gauche, l'axe 1 est celui de la corrélation avec un maximum de variables, en particulier les principales matières. Le n° 8 est la tête de classe, le groupe à l'opposé réunit 18, 14, 21, 19, 4 et 11. On peut vérifier sur le tableau de données.

16.

Sur la carte factorielle des lignes, les valeurs négatives sur le deuxième axe sont liées à des valeurs élevées en anglais et espagnol (carte des variables en haut à droite. Le n° 12 est assez bon mais faible en langues. Le n° 11 est plutôt faible mais bon en langues. On vérifie sur les données :

	HGEO	FRAN	PHYS	MATH	BIOL	ECON	ANGL	ESPA	
11	*10.6*	8.0	*4.5*	*7.1*	*9.8*	*8.0*	**14**	**16.5**	
12	**13.4**	8.8	**13.0**	**13.8**	**12.6**	10.0	*9*	*11.5*	
m	12.2	10.1	8.5	9.3	11.2	11.2	10.2	13.9	Les moyennes servent de référence

17.

```
L'inertie totale d'une ACO normée égale le nombre de variables.
> sum(pca1$eig)
[1] 8
Les axes principaux sont orthonormés pour la métrique canonique
> t(pca1$c1)%*%as.matrix(pca1$c1)

             CS1       CS2
  CS1 1.000e+00 2.087e-18
  CS2 2.087e-18 1.000e+00
Les coordonnées factorielles sont de variance lambda et non covariantes
> t(pca1$li)%*%as.matrix(pca1$li)/22

              Axis1      Axis2
  Axis1  3.493e+00 -1.741e-16
  Axis2 -1.741e-16  1.491e+00
L'axe principal a des composantes proportionnelles aux coordonnnées des variables.
> (pca1$co[1,1]/sqrt(pca1$eig[1])) == pca1$c1[1,1]
[1] TRUE
```

18.

```
> eigen(cor(pca1$tab))$values[2]
[1] 1.4911
Les valeurs propres de la matrice de corrélation sont celles de l'ACP du tableau
> eigen(pca1$tab)$values[2]
Error in eigen(pca1$tab) : non-square matrix in eigen
Le message d'erreur indique que la diagonalisation d'une matrice non carrée n'a pas de
sens.
```

19.

```
> svd(cor(pca1$tab))$d[2]
[1] 1.4911
Les valeurs singulières de R sont les racines carrées des valeurs propres de RR car c'est
une matrice symétrique, donc les racines carrées des carrés des valeurs propres, donc les
valeurs propres de l'ACP
```

20.

```
> svd(pca1$tab/sqrt(22))$d[2]^2
[1] 1.4911
```

Les carrés des valeurs singulières de $\mathbf{X}/\sqrt{n}$ sont les valeurs propres de $\mathbf{X}^t\mathbf{X}/\sqrt{n}\sqrt{n} = (1/n)\mathbf{X}^t\mathbf{X} = \mathbf{R}$, donc les valeurs propres de l'ACP.
