Introduction à SPlus

Exercice 1

0- Les résumés unidimensionnels :

```
> Y_c(4,1,0,3,6,3,0,3,1,4,3,0,3,6,3,0,1,1,3,3,0,0,6,6)
> dim(Y)_c(8,3)
> summary(data.frame(Y))

       Y.1             Y.2             Y.3
    Min.:0.00       Min.:0.00       Min.:0.00
 1st Qu.:0.75    1st Qu.:0.75    1st Qu.:0.75
  Median:3.00     Median:3.00     Median:2.00
       Y.1             Y.2             Y.3
    Mean:2.50       Mean:2.50       Mean:2.50
 3rd Qu.:3.25    3rd Qu.:3.25    3rd Qu.:3.75
    Max.:6.00       Max.:6.00       Max.:6.00
```

1- Les moyennes des variables colonnes

```
> weighted.mean(Y)
[1] 2.5
```

2- Le tableau centré

On retranche sa moyenne à chacune des colonnes :

```
> centre(Y)
     [,1] [,2] [,3]
[1,]  1.5 -1.5 -1.5
[2,] -1.5  1.5 -1.5
[3,] -2.5  0.5  0.5
[4,]  0.5 -2.5  0.5
[5,]  3.5  0.5 -2.5
[6,]  0.5  3.5 -2.5
[7,] -2.5  0.5  3.5
[8,]  0.5 -2.5  3.5
```

3- La matrice de covariance

```
> varx(Y)
      [,1]  [,2]  [,3]
[1,]  3.75 -0.75 -2.25
[2,] -0.75  3.75 -2.25
[3,] -2.25 -2.25  5.25
```

4- La matrice de corrélation

On obtient le vecteur des corrélations et des écarts-types sigma :

```
> corx(Y)
           [,1]       [,2]       [,3]
[1,]  1.0000000 -0.2000000 -0.5070926
[2,] -0.2000000  1.0000000 -0.5070926
[3,] -0.5070926 -0.5070926  1.0000000
attr(, "sigma"):
[1] 1.936492 1.936492 2.291288
```

5- Les variances et écarts-types

Variance :

```
> diag(varx(Y))

[1] 3.75 3.75 5.25
```

Écarts-Types :

On retrouve les memes resultats qu'au sigma du (4-)

```
> sqrt(diag(varx(Y)))

[1] 1.936492 1.936492 2.291288
```

Exercice 2

1) Nombre de mots par ligne

On réalise 13 observations pondérées, puis on reconstitue le tableau des données :

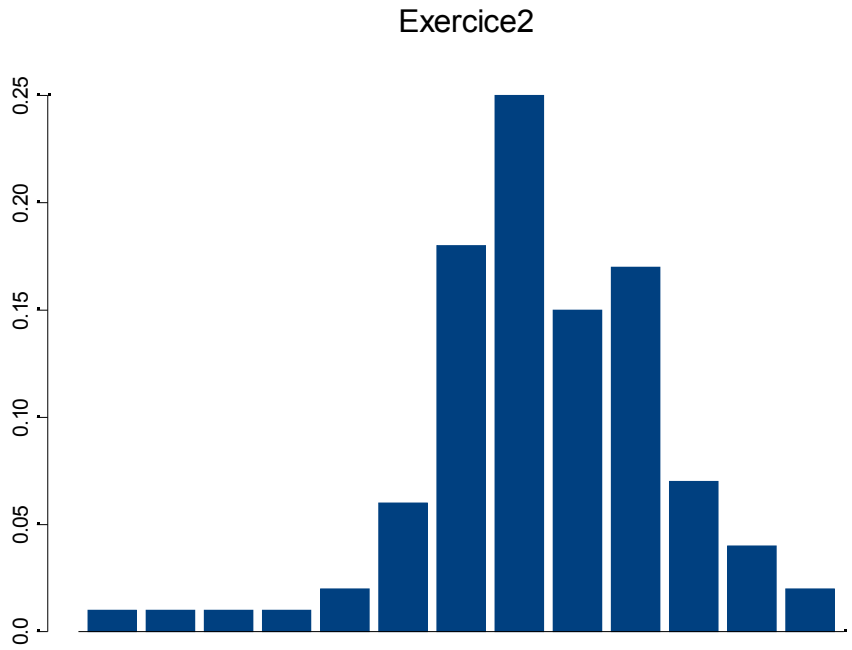
```
> X_c(5,6,7,8,9,10,11,12,13,14,15,16,17) #matrice des données
> P_c(1/100,1/100,1/100,1/100,1/50,3/50,9/50,1/4,3/20,17/100,7/100,1/25,1/50)
#poids
> Y_rep(X,100*P)
> Y
```

```
 [1]  5  6  7  8  9  9 10 10 10 10 10 10 11 11 11 11 11 11 11 11 11 11 11 11 11
  11 11 11 11 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12
  12 12 13 13 13
[59] 13 13 13 13 13 13 13 13 13 13 13 13 14 14 14 14 14 14 14 14 14 14 14 14 14
  14 14 14 15 15 15 15 15 15 15 16 16 16 16 17 17
```

```
> dim(Y)_c(100,1)
> varx(Y)

       [,1]
[1,] 4.3156

> sqrt(varx(Y))

         [,1]
[1,] 2.077402
```

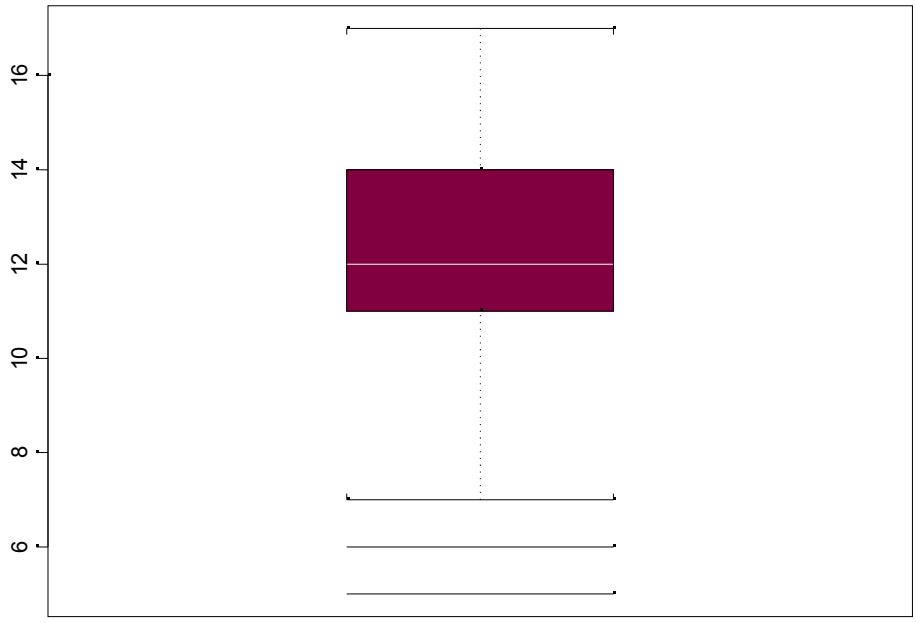
Affichage des diagrammes :

```
> v_1:13
> abs_v+.2*(v-2)
> X11()
> barplot(P)
> text(abs,-1,X)
> title("Exercice2")
```



On observe que le nombre moyen de mots par ligne est d'environ 12 et que la plupart des lignes a entre 10 et 14 mots.

```
> boxplot(Y)
```



On retrouve les mêmes résultats qu'avec le barplot mais en un peu plus lisible. Ce graphique fournit plusieurs informations intéressantes sur le tableau que nous avons entré :
Le bloxplot nous donne la valeur de la médiane (environ 12), ainsi que les différents quartiles. En noir sont représentés les deuxièmes et troisièmes quartiles. Le premier quartile nous informe que 25% des valeurs se trouvent entre 7 et 11, si on ne tient pas compte des valeurs originales. En effet, celles-ci sont représentées par des traits horizontaux à l'extérieur du boxplot. Cette représentation est destinée à ne pas fausser la visualisation du contexte général à cause de quelques observations originales.

2) Longueur des feuilles de platane

On étudie 75 feuilles de platane réparties en classes de longueur.

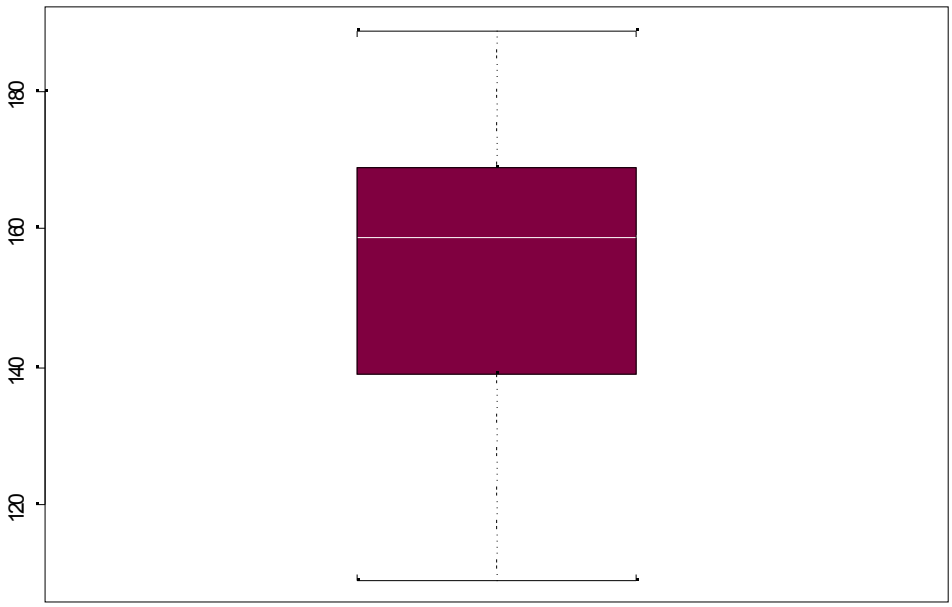
```
> X_c(104,109,114,119,124,129,134,139,144,149,154,159,164,169,174,179,184,189)
#classes
> N_c(1,0,3,3,2,4,5,4,6,9,10,7,6,4,5,3,2,1) #nombre d'éléments par classe
> dim(X)_c(18,1)
> X
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]  104  109  114  119  124  129  134  139  144   149

     [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18]
[1,]   154   159   164   169   174   179   184   189

> grav(X,N)

[1] 150.3333

> Y_rep(X,N)
> dim(Y)_c(75,1)
> varx(Y)
          [,1]
[1,] 354.2222

> sqrt(varx(Y))
          [,1]
[1,] 18.82079
```

Si on regroupe les classes par 2 :

```
> X_c(109,119,129,139,149,159,169,179,189)
> N_c(1,6,6,9,15,17,10,8,3)
> dim(X)_c(9,1)
> grav(X,N)


[1] 153

> Y_rep(X,N)
> dim(Y)_c(75,1)
> varx(Y)

           [,1]
[1,] 357.3333

> sqrt(varx(Y))
         [,1]
[1,] 18.90326
```

On observe que les résultats sont légèrement supérieurs à ceux obtenus précédemment. En effet il aurait été plus judicieux de faire la moyenne pondérée des classes lors de la fusion desdites classes.

```
> boxplot(Y)
```



Indices statistiques de base

Exercice 1

Exemple 1 :

```
> exp1_c(-4,-1,-1,-4,-1,-1,2,4,4,2,2,3,1,-2,-3,-1,1,1,-1,-1)
> dim(exp1)_c(10,2)
> P1_c(1,1,1,2,2,2,3,3,3,3) #nous permet de definir 3 classes dans exp1
> exp1
      [,1] [,2]
 [1,]   -4    2
 [2,]   -1    3
 [3,]   -1    1
 [4,]   -4   -2
 [5,]   -1   -3
 [6,]   -1   -1
 [7,]    2    1
       [,1] [,2]
  [8,]    4    1
  [9,]    4   -1
 [10,]    2   -1
```

Calcul de l'inertie totale :

```
> inerx(exp1)

[1] 10.8
```

Calcul de l'inertie inter-classes :

```
> interx(exp1,P1)


[1] 8.4
```

La formule intertie totale = inertie inter + inertie intra nous donne :
inertie intra = 2.4

Même calcul à partir du carré des distances inter-points :

```
> d_dist(exp1) #distances inter-point
> inerd(d) #distance euclidienne par defaut

[1] 10.8
```

On continue de même pour chacune des classes :

```
> i1_inerd(dist(exp1[1 :3,]))
> i2_inerd(dist(exp1[4 :6,]))
> i3_inerd(dist(exp1[7 :10,]))
> 0.3*i1+0.3*i2+0.4*i3 #inertie intra


[1] 2.4
```

On retrouve les mêmes résultats que précédemment.

Exemple 2

```
> exp2_c(4,1,0,3,6,3,0,3,1,4,3,0,3,6,3,0,1,1,3,3,0,0,6,6)
> dim(exp2)_c(8,3)
     [,1] [,2] [,3]
[1,]    4    1    1
[2,]    1    4    1
[3,]    0    3    3
[4,]    3    0    3
[5,]    6    3    0
[6,]    3    6    0
[7,]    0    3    6
[8,]    3    0    6
```

```
> C2_c(1,1,1,1,2,2,2,2) #permet de definir les classes et les poids
```

Calcul des poids et des centres de gravités des classes.

```
> poidsc(C2,C2/12)

[1] 0.3333333 0.6666667

> gravc(exp2,C2,C2/12)
  [,1] [,2] [,3]
1    2    2    2
2    3    3    3
attr(, "poidscl"):
[1] 0.3333333 0.6666667
```

Calcul des variances totales, inter et intra :

```
> varx(exp2,C2/12) #matrice de covariance

           [,1]       [,2]      [,3]
[1,]  4.0555556 -0.4444444 -2.944444
[2,] -0.4444444  4.0555556 -2.944444
[3,] -2.9444444 -2.9444444  6.555556

> between(exp2,C2,C2/12) #covariance inter


          [,1]      [,2]      [,3]
[1,] 0.2222222 0.2222222 0.2222222
[2,] 0.2222222 0.2222222 0.2222222
[3,] 0.2222222 0.2222222 0.2222222

> within(exp2,C2,C2/12) #covariance intra


           [,1]       [,2]      [,3]
[1,]  3.8333333 -0.6666667 -3.166667
[2,] -0.6666667  3.8333333 -3.166667
[3,] -3.1666667 -3.1666667  6.333333
```

Calcul des inerties totales et inter

```
> inerx(exp2,C2/12) #totale


[1] 14.66667

> interx(exp2,C2,C2/12) #inter


[1] 0.6666667
```

Les centres de gravités des classes sont très proches car la covariance inter est petite devant la covariance intra.

Analyse en composantes principales

Exercice 1 : ACP Notes

```
> notes_c
(6,8,6,14.5,14,11,5.5,13,9,6,8,7,14.5,14,10,7,12.5,9.5,5,8,11,15.5,12,5.
5,14,8.5,12.5,5.5,8,9.5,15,12.5,7,11.5,9.5,12,8,9,11,8,10,13,10,12,18)
> dim(notes)_c(9,5)
> dimnames(notes)_list(c("JE", "AL", "AN", "MO", "DI", "AD", "PI", "BR",
"EV"), c("m", "s", "f", "l", "dm"))
```

ACP sur les variables centrées réduites :

```
> res_acp(notes)
```

ACP du tableau "notes" sur variables reduites

Inertie totale : 5

Inertie expliquee (en %):

```
      f1 f2  f3  f4  f5
      58 23  20   0   0
Cumul 58 80 100 100 100
```

Composantes principales :

```
> res$vectors
```

```
           f1         f2          f3          f4         f5
 m 0.47764775 -0.5287000 -0.16117378 -0.29869388  0.6141123
 s 0.53101865 -0.3959114 -0.09874597  0.52660628 -0.5236491
 f 0.44395376  0.5779663  0.23195987  0.46927256  0.4414045
 l 0.53983875  0.3604182  0.11348528 -0.64195022 -0.3920381
dm 0.03675592  0.3158865 -0.94741119  0.03392601  0.0112169
```

Les valeurs propres correspondantes :

```
> res$values
```

```
       f1       f2       f3          f4           f5
 2.878226 1.134815 0.983605 0.002373772 0.0009810042
```

L'inertie nous montre que les 3 premières composantes principales conservent la totalité de l'inertie et donc concentrent la majeure partie de l'information.

Nouveau tableau de données dans la base des composantes principales :

```
> reconacp(res)
```

```
      m    s    f    l dm
JE  6.0  6.0  5.0  5.5  8
AL  8.0  8.0  8.0  8.0  9
AN  6.0  7.0 11.0  9.5 11
MO 14.5 14.5 15.5 15.0  8
DI 14.0 14.0 12.0 12.5 10
AD 11.0 10.0  5.5  7.0 13
PI  5.5  7.0 14.0 11.5 10
BR 13.0 12.5  8.5  9.5 12
EV  9.0  9.5 12.5 12.0 18
```

```
> contri(res)
Contribution des axes principaux aux individus

    Axe1 Axe2 Axe3 Axe4 Axe5  Tot
 JE  883   55   62    0    0 1000
 AL  790   58  151    0    0 1000
 AN  489  469   41    2    0 1000
 MO  879    3  117    0    0 1000
 DI  879  112    8    0    1 1000
 AD  246  390  363    0    1 1000
 PI   30  755  215    0    0 1000
 BR  183  584  230    3    0 1000
 EV   54  353  592    0    0 1000
Tot  576  227  197    0    0 1000


Qualite de la representation des individus
 sur les ss-esp. principaux

    Axe 1 Axes1a2 Axes1a3 Axes1a4 Axes1a5  tot
 JE   883     938    1000    1000    1000 1000
 AL   790     849     999    1000    1000 1000
 AN   489     957     998    1000    1000 1000
 MO   879     882    1000    1000    1000 1000
 DI   879     991     999     999    1000 1000
 AD   246     636     999     999    1000 1000
 PI    30     785    1000    1000    1000 1000
 BR   183     767     997    1000    1000 1000
 EV    54     407     999    1000    1000 1000
Tot   576     803     999    1000    1000 1000

(on voit bien que les 2 premiers axes suffisent à décrire assez bien les
individus sauf dans le cas de l'individu EV)

Contribution des individus aux axes principaux

    Axe1 Axe2 Axe3 Axe4 Axe5  Tot  ORI
 JE  298   47   61  122  150  194  194
 AL   62   12   34   41   23   45   45
 AN   41   99   10  155    3   48   48
 MO  371    3  145  146  121  243  243
 DI  160   52    4   21  401  105  105
 AD   35  140  150    2  247   81   81
 PI    5  294   97   72   19   89   89
 BR   15  124   56  300    0   48   48
 EV   14  229  443  141   36  147  147
Tot 1000 1000 1000 1000 1000 1000 1000


Contribution des facteurs aux variables

     f1  f2  f3 f4 f5  tot
  m 657 317  26  0  0 1000
  s 812 178  10  1  0 1000
  f 567 379  53  1  0 1000
  l 839 147  13  1  0 1000
 dm   4 113 883  0  0 1000
Tot 576 227 197  0  0 1000


Qualite de la representation des variables
 sur les ss-esp. principaux
```

	f1	f1a2	f1a3	f1a4	f1a5	tot
m	657	974	999	1000	1000	1000
s	812	989	999	1000	1000	1000
f	567	946	999	1000	1000	1000
l	839	986	999	1000	1000	1000
dm	4	117	1000	1000	1000	1000
Tot	576	803	999	1000	1000	1000

(les 2 premières composantes donnent une bonne représentation des variables sauf pour dm)

Contribution des variables aux facteurs

	f1	f2	f3	f4	f5	Tot	VARI
m	228	280	26	89	377	200	233
s	282	157	10	277	274	200	183
f	197	334	54	220	195	200	246
l	291	130	13	412	154	200	162
dm	1	100	898	1	0	200	177
	1000	1000	1000	1000	1000	1000	1000

A partir des résultats on déduit facilement que les deux premières composantes principales suffisent à conserver l’essentiel des informations