MAITRISE BPE - UV MIAB 2 - JUIN 2004

STATISTIQUES (2 HEURES)

Pour une question marquée par **?** indiquer si l'assertion est juste ou fausse et justifier par un argument qui vous semble solide. Il est inutile de répondre à toutes les questions : on pourra obtenir un bon résultat en choisissant son point de vue. La première partie est plus algébrique, la seconde plus statistique et la troisième plus technique. Merci de répondre dans les cadres impartis à cet effet. *Tous les documents sont autorisés*.

PARTIE I

1. **?** Pour deux variables statistiques centrées et de variances non nulles, le nuage des points est sur une droite si et seulement si le coefficient de corrélation égale à 1.

2. **?** Le pourcentage de variance expliquée par la prédiction linéaire de **x** par **y** est toujours égal au pourcentage de variance expliquée par la prédiction linéaire de **y** par **x**.

3. **?** Toute matrice carrée de nombre réels a au moins un vecteur propre.

4. **?** Toute matrice carrée de nombre réels a au moins un vecteur non nul qui n'est pas propre.

5. **?** Toute matrice carrée de nombre réels qui a des vecteurs propres est diagonalisable.

6. **?** Les valeurs propres d'une matrice de corrélation sont toujours positives ou nulles.

7. **?** Un tableau **X** a *n* lignes-individus et *p* colonnes-variables et **C** est sa matrice de covariances. Si la somme par lignes dans **X** vaut 1 pour toutes les lignes, 0 est toujours valeur propre de **C**.

8. **?** Le rang de la matrice de corrélation d'une ACP normée sur *p* variables vaut toujours *p*.

9. **?** Dans une matrice de corrélation théorique sur p variables dans laquelle tous les coefficients non diagonaux sont égaux à α une valeur propre vaut $1+(p-1)\alpha$.

10. **?** L'inertie projetée sur le premier axe principal d'une ACP normée sur *p* variables vaut au moins 1.

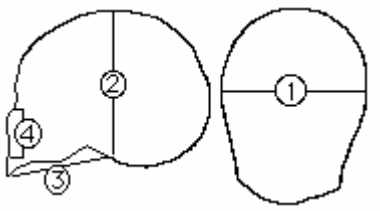
PARTIE II

Les données utiles, dans le data.frame **`skulls`**, sont extraites de l'ouvrage de Manly, B.F. (1994) *Multivariate Statistical Methods. A primer*. Second edition. Chapman & Hall, London.

1-215. Les mesures concernent 5 groupes de 30 crânes égyptiens. Les classes sont les modalités d'un facteur **fac** à 5 niveaux :

1 - période prédynastique ancienne (4000 avant JC)
2 - période prédynastique récente (3300 avant JC)
3 - 12 ème et 13 ème dynasties (1850 avant JC)
4 - période de Ptolémée (200 avant JC)
5 - période romaine (150 après JC)

Les variables ont des distributions sensiblement normales et sont définies par :




```
apply(skulls,2,function(x) anova(lm(x~fac)))
$V1
Analysis of Variance Table

Response: x
           Df Sum Sq Mean Sq F value  Pr(>F)
fac         4    503     126    5.95 0.00018
Residuals 145   3061      21

$V2
Analysis of Variance Table

Response: x
           Df Sum Sq Mean Sq F value Pr(>F)
fac         4    230      57    2.45  0.049
Residuals 145   3405      23

$V3
Analysis of Variance Table

Response: x
           Df Sum Sq Mean Sq F value  Pr(>F)
fac         4    803     201    8.31 4.6e-06
Residuals 145   3506      24

$V4
Analysis of Variance Table

Response: x
           Df Sum Sq Mean Sq F value Pr(>F)
fac         4     61      15    1.51   0.20
Residuals 145   1472      10
```

11. Au vu de ce qui précède, peut-on parler d'évolution des mesures au cours de la période analysée ?

12. Qu'attendez-vous du résultat des deux lignes de commande qui suivent ?

```
facord=as.ordered(fac)
apply(skulls,2,function(x) anova(lm(x~facord)))
```

On donne en outre :

```
summary(lm(skulls$V3~fac))
Call:
lm(formula = skulls$V3 ~ fac)
```

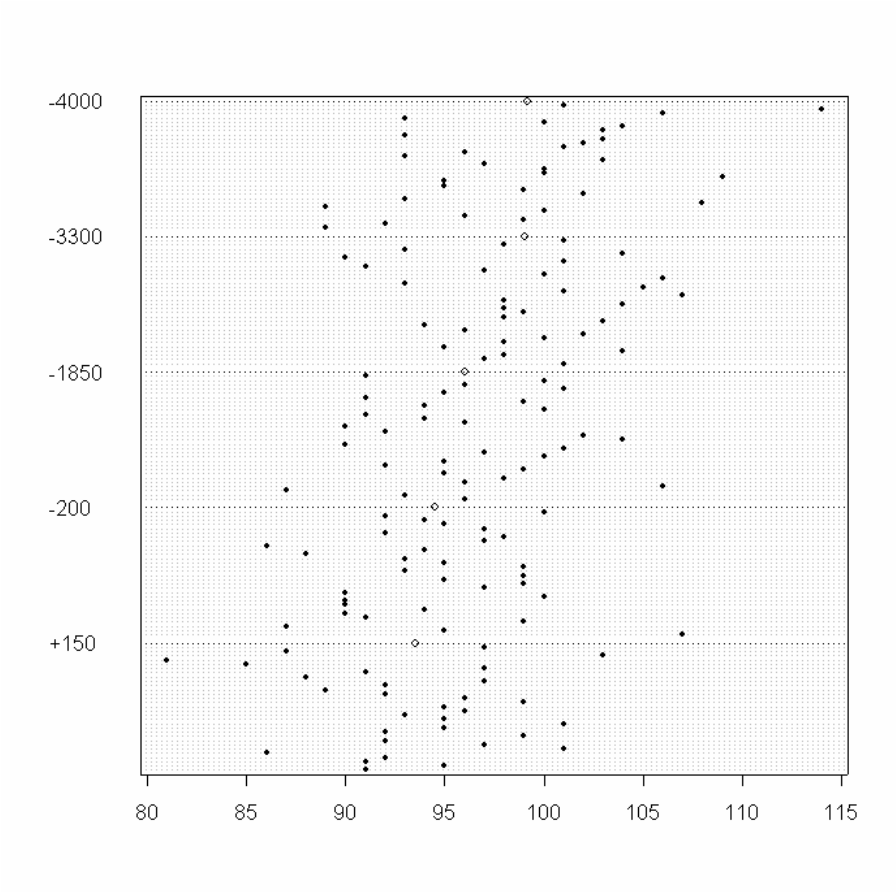
```
Residuals:
     Min       1Q   Median       3Q      Max
-12.5000  -3.1667  -0.0333   3.5000  14.8333

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   99.167      0.898  110.46  < 2e-16
fac-3300      -0.100      1.270   -0.08  0.93733
fac-1850      -3.133      1.270   -2.47  0.01475
fac-200       -4.633      1.270   -3.65  0.00037
fac+150       -5.667      1.270   -4.46  1.6e-05

Residual standard error: 4.92 on 145 degrees of freedom
Multiple R-Squared: 0.186,      Adjusted R-squared: 0.164
F-statistic: 8.31 on 4 and 145 DF,  p-value: 4.64e-06
```

```
summary(lm(skulls$V3~facord))
Call:
lm(formula = skulls$V3 ~ facord)

Residuals:
     Min       1Q   Median       3Q      Max
-12.5000  -3.1667  -0.0333   3.5000  14.8333

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  96.4600     0.4015  240.26  < 2e-16
facord.L     -5.0175     0.8978   -5.59  1.1e-07
facord.Q     -0.0891     0.8978   -0.10     0.92
facord.C      1.0752     0.8978    1.20     0.23
facord^4     -0.6614     0.8978   -0.74     0.46

Residual standard error: 4.92 on 145 degrees of freedom
Multiple R-Squared: 0.186,      Adjusted R-squared: 0.164
F-statistic: 8.31 on 4 and 145 DF,  p-value: 4.64e-06
```

13. Ces deux résultats sont-ils cohérents ? Si oui quelle information commune expriment-ils? Si non, d'où vient la contradiction ?

On fait :

```
dotchart(skulls$V3, gr=fac,gdata=tapply(skulls$V3,fac,mean),pch=20)
```



14. Quelles sont et que signifient les valeurs numériques représentées par des petits cercles ?

On fait l'ACP normée de ce tableau :

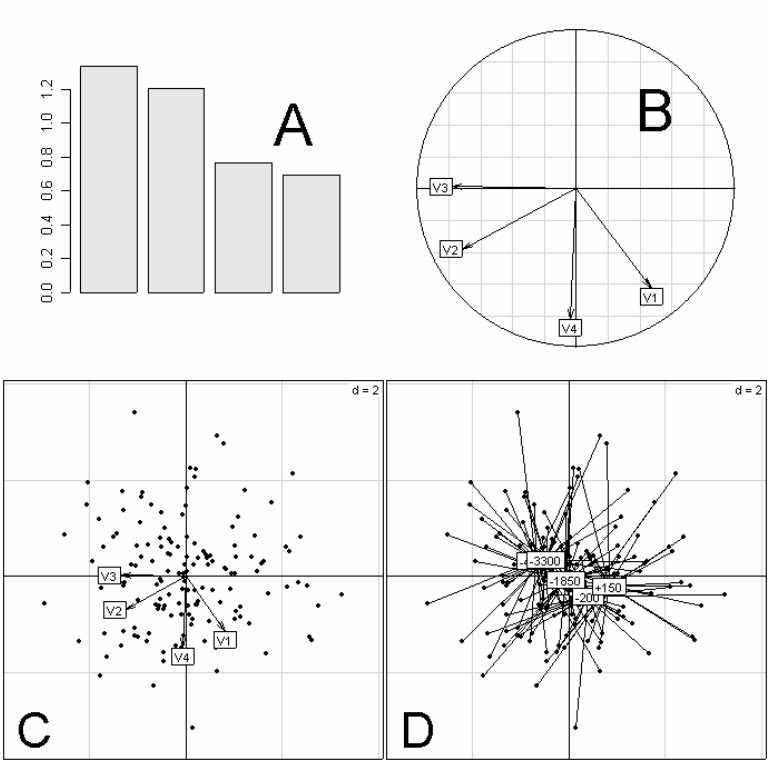
```
pca1=dudi.pca(skulls)
Select the number of axes: 2

pca1$eig
[1] 1.3373 1.2064 0.7624 [redacted]
```

15. Quelle est la dernière valeur propre ? Quel est le pourcentage de variance représenté sur les deux premiers axes ?

16. Quelle relation lie les deux matrices éditées ci-dessous ?

```
cor(skulls)
        V1      V2       V3       V4
V1  1.0000 -0.0619 -0.15697  0.18255
V2 -0.0619  1.0000  0.26435  0.14675
V3 -0.1570  0.2644  1.00000 -0.00638
V4  0.1825  0.1468 -0.00638  1.00000

 pca1$c1
        CS1      CS2
V1  0.40698 -0.56741
V2 -0.61718 -0.34501
V3 -0.67245  0.01276
V4 -0.03554 -0.74756

par(mfrow=c(2,2))
barplot(pca1$eig,col=grey(0.9)) ; text(4,1,"A",cex=4)
s.corcircle(pca1$co) ; text(0.5,0.5,"B",cex=4)
s.label(pca1$li, cpoi=1, clab=0)
```

```
s.arrow(2*pca1$c1, add.p=T,sub="C",csub=4)
s.class(pca1$li,fac,cell=0,sub="D",csub=4
```



17. Donner une légende technique à cette figure.

18. Quelle relation voyez-vous entre la matrice calculée par `cor(skulls)` et le graphe ci-dessus ?

19. Donner approximativement la corrélation entre la coordonnée des individus sur le premier axe et chacune des variables.

20. Que faut-il faire pour centrer l'analyse sur la différence entre les périodes ?

PARTIE III

Deux relevés extraits d'un tableau faunistique sont :

```
  CHA TRU VAI LOC OMB BLA HOT TOX VAN CHE BAR SPI GOU BRO PER BOU PSO ROT CAR
18   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   0   1
26   0   0   0   1   0   0   1   0   1   1   1   1   1   1   1   1   1   1
   TAN BCO PCH GRE GAR BBO ABL ANG
18   1   0   0   1   1   0   1   1
26   1   1   1   1   1   1   1   1
```

21. Quelle est leur dissimilarité mesurée au sens de l'indice de communauté de Jaccard ?

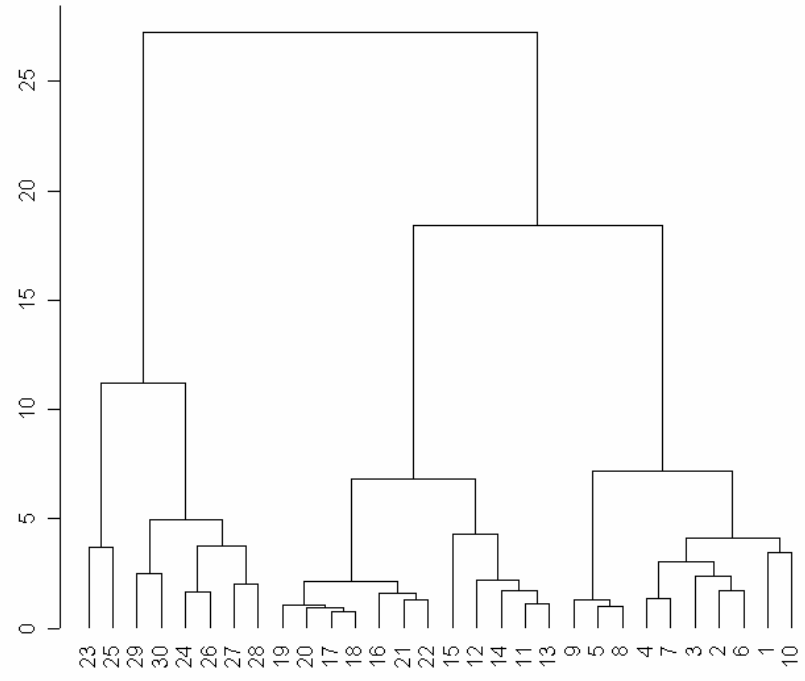
22. Qu'est-ce qu'une matrice de distance ultramétrique ?

Dans le cours, on rencontre plusieurs fois le tableau doubs$poi à 30 stations et 27 espèces. Le tableau doubs$mil contient des mesures de 11 variables environnementales sur les mêmes 30 stations du Doubs :

```
doubs$mil
    das alt   pen  deb pH dur pho nit amm oxy dbo
1     3 934 6.176   84 79  45   1  20   0 122  27
2    22 932 3.434  100 80  40   2  20  10 103  19
3   102 914 3.638  180 83  52   5  22   5 105  35
...
27 3732 206 2.565 3960 81  90  58 300  26  72  63
28 3947 195 1.386 4320 83 100  74 400  30  81  45
29 4220 183 1.946 6770 78 110  45 162  10  90  42
30 4530 172 1.099 6900 82 109  65 160  10  82  44

das = distance à la source en m
alt = altitude en m
pen = pente en pour pour mille
deb = débit en m3/s
pH = potentiel Hydrogène
dur = dureté calcique en mg/l
pho = phosphates en mg/l
nit = nitrates en mg/l
amm = azote en mg/l
oxy = oxygène dissous en % de saturation
dbo = demande biologique en oxygène en mg/l
```

23. Donner une légende technique à la figure ci-dessous et obtenue par :

```
tab0=scalewt(doubs$mil)
h0=hclust(dist(tab0),"ward")
plot(h0,hang=-1)
```



24. Quel sera le résultat de l'opération suivante et à quoi sert-elle ?

```
cutree(h0,3)
```

25. Donner quelques idées pour continuer l'analyse de ces données.

MAITRISE BPE - UV MIAB 2 - JUIN 2004 - STATISTIQUES (2 HEURES)

1. **?** Pour deux variables statistiques centrées …

Non. Contre-exemple : il l'est aussi si r=-1

2. **?** Le pourcentage de variance expliquée …

Oui, car dans les deux cas c'est le carré du cosinus de l'angle des deux vecteurs centrés.

3. **?** Toute matrice carrée de nombre réels a au moins un vecteur propre.

Non. Contre-exemple : la matrice d'une rotation d'angle 45° (par exemple !) n'a aucun vecteur propre, parce qu'un vecteur propre ne peut être nul et qu'un vecteur non nul ne peut être proportionnel à lui-même après rotation.

4. **?** Toute matrice carrée de nombre réels a au moins un vecteur non nul qui n'est pas propre.

Non. Contre-exemple, la matrice identité vérifie pour tout vecteur $\mathbf{Iu} = \mathbf{u} = 1\mathbf{u}$ et tout vecteur est propre pour 1.

5. **?** Toute matrice carrée de nombre réels qui a des vecteurs propres est diagonalisable.

Non. Contre-exemple, la matrice $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ a un vecteur propre $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ pour la valeur propre 0 mais on ne peut pas trouver une base de vecteurs propres. En effet :

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \lambda\begin{bmatrix} x \\ y \end{bmatrix} \Rightarrow \begin{cases} y = \lambda x \\ 0 = \lambda y \end{cases} \Rightarrow \begin{cases} \lambda = 0 \Rightarrow y = 0 \\ ou \\ \lambda \neq 0 \Rightarrow y = 0 \Rightarrow x = 0 \end{cases}$$

6. **?** Les valeurs propres d'une matrice de corrélation sont toujours positives ou nulles.

Oui, car une matrice de corrélation, à partir du tableau centré s'écrit $\mathbf{R} = \frac{1}{n}\mathbf{Y}^t\mathbf{Y}$. Alors :

$$\mathbf{Ru} = \lambda\mathbf{u} \Rightarrow \frac{1}{n}\mathbf{Y}^t\mathbf{Yu} = \lambda\mathbf{u} \Rightarrow \mathbf{Y}^t\mathbf{Yu} = \lambda n\mathbf{u} \Rightarrow \mathbf{u}^t\mathbf{Y}^t\mathbf{Yu} = \lambda n\mathbf{u}^t\mathbf{u}$$

$$\Rightarrow \lambda = \frac{\mathbf{u}^t\mathbf{Y}^t\mathbf{Yu}}{n\mathbf{u}^t\mathbf{u}} = \frac{1}{n}\frac{\|\mathbf{Yu}\|^2}{\|\mathbf{u}\|^2}$$

7. **?** Un tableau **X** a *n* lignes-individus et *p* colonnes-variables et **C** est sa matrice …

Oui, car si **Y** est le tableau centré associé à **X** on a :

$$\mathbf{X1}_p = 1 \Rightarrow \mathbf{Y1}_p = 0 \Rightarrow \frac{1}{n}\mathbf{Y}^t\mathbf{Y1}_p = \mathbf{C1}_p = 0 \Rightarrow \mathbf{1}_p \text{ propre pour } 0$$

8. **?** Le rang de la matrice de corrélation d'une ACP normée sur *p* variables vaut toujours *p*.

Non, contre-exemple, le tableau à 2 colonnes, formé de deux fois la même variable, a une matrice de corrélation qui vaut $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, laquelle est de rang 1.

9. **?** Dans une matrice de corrélation théorique sur p variables dans laquelle …

Oui, car :

$$\begin{bmatrix} 1 & \alpha & \cdots & \alpha & \alpha \\ \alpha & 1 & \cdots & \alpha & \alpha \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \alpha & \alpha & \cdots & 1 & \alpha \\ \alpha & \alpha & \cdots & \alpha & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix} = \left(1+\left(p-1\right)\alpha\right) \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix}$$

10. **?** L'inertie projetée sur le premier axe principal d'une ACP normée sur p variables …

Oui, car l'inertie totale vaut la trace de la matrice de corrélation, donc le nombre de variable. Elle vaut aussi la somme des valeurs propres. Si la plus grande valeur propre, qui est aussi l'inertie projetée sur le premier axe, était inférieure à 1, la somme serait inférieure à p, ce qui est contradictoire.

11. Peut-on parler d'évolution des mesures au cours de la période analysée ?

Non, car on ne sait rien des moyennes par classes. On peut simplement dire que l'hypothèse nulle de l'égalité des moyennes par échantillons est presque sûrement fausse pour les variables 1 et 3, douteuse pour la variable 2 et qu'on n'a aucun argument pour la rejeter dans le cas de la variable 4.

12. Qu'attendez-vous du résultat des deux lignes de commande qui suivent ?

Exactement le même résultat que pour les anova utilisant le facteur fac à modalités non ordonnées, car on n'a strictement rien changé au sous-espace de projection.

13. Ces deux résultats sont-ils cohérents ? …

Ces deux résultats sont complètement cohérents. Ils expriment de deux manières très différentes que la variable V3 en moyenne diminue avec le temps. Avec le facteur, les contrastes par défaut placent les témoins dans la classe 1 et donc affectent à la première classe la valeur 0. Les valeurs suivantes -0.10, -3.13, -4.6, -5.67 sont décroissantes et la dernière est très significativement négative. Avec les modalités ordonnées, le contraste linéaire est très significatif et la pente de -5.01 indique la décroissance. Les anova dans les deux cas sont les mêmes.

14. Quelles sont et que signifient les valeurs numériques représentées par des petits cercles ?

Ce sont les moyennes par classes et elles valent :

```
99.17, 99.17-0.1 = 99.07, 99.17-3.13 = 96.04,

99.17 - 4.63 =  94.54 et 99.17 - 5.67 = 93.50
```

15. Quelle est la dernière valeur propre ? …

```
La somme fait 4, donc elle vaut 0.6939.

 (1.3373 + 1.2064) / 4 = 0.6359
```

16. Quelle relation lient les deux matrices éditées ci-dessous ?

Si on appelle **R** la première et **B** le seconde, comme il s'agit de la matrice de corrélation et de la matrice des deux premiers axes principaux, on a :

$$\mathbf{RB} = \mathbf{B}\begin{bmatrix} 1.3373 & 0 \\ 0 & 1.2064 \end{bmatrix}$$

17. Donner une légende technique à cette figure.

Représentations graphiques associées à une analyse en composantes principales normée d'un tableau de 150 lignes individus réparties en 5 groupes et de 4 variables morphométriques. A : graphe des valeurs propres donnant la répartition de l'inertie entre les axes principaux. B : cercle des corrélations comme projection des quatre variables normées sur le plan des deux premières composantes principales. Le cercle est l'intersection du plan avec la sphère unité de $\mathbb{R}^{150}$. C : carte factorielle des individus ou projection des 150 lignes du tableau normé sur les 2 premiers axes principaux. D : le même plan est utilisé pour représenter les cinq groupes par des étoiles reliant chaque point au centre de gravité du groupe auquel il appartient.

18. Quelle relation voyez-vous entre la matrice calculée par `cor(skulls)` et le graphe ci-dessus ?

La matrice de corrélation donne de faibles corrélations, donc des variables en grande partie indépendantes. On retrouve cela sur le graphe des valeurs propres relativement plat. V1 et V2 sont indépendantes, V3 et V4 aussi et les vecteurs sont presque orthogonaux. L'axe 1 prend en compte la faible corrélation V2-V3 (`0.264`) et l'axe 2 prend en compte la faible corrélation V1-V4 (`0.183`).

19. Donner approximativement la corrélation entre …

A lire sur le cercle des corrélations, environ `0.5, -0.7, -0.8 et 0`, avec plus de précision `(pca1$co[,1]) : 0.4706 -0.7137 -0.7776 -0.0411`

20. Que faut-il faire pour centrer l'analyse sur la différence entre les périodes ?

L'analyse en composantes principales simple du tableau des moyennes par classes des variables normalisée (ACP inter-classes) ou mieux la recherche des combinaisons linéaires de variables de variance unité maximisant la variance inter-classe (Analyse discriminante).

21. Quelle est leur dissimilarité mesurée au sens de l'indice de communauté de jaccard ?

```
a = 17, b = 6, c = 4, d = 0 S = 17/27, d = sqrt(10/27) = 0.6086
```

22. Qu'est-ce qu'une matrice de distance ultramétrique ?

C'est une matrice de distances telle que, pour trois individus arbitraires x, y et z, on a toujours $d(x,y) \leq \max\left(d(x,z), d(x,z)\right)$.

23. Donner une légende technique à la figure ci-dessous et obtenue par :

Dendrogramme ou représentation graphique d'une classification ascendante hiérarchique. Le tableau donnant 11 variables pour 30 stations est normalisée par variable (scale). On calcule la distance euclidienne entre lignes (dist) sur le tableau normalisé, c'est-à-dire la distance

canonique en usage dans l'ACP normée du tableau. Les individus puis les parties sont regroupées dans une hiérarchie de parties en faisant à chaque étape le regroupement qui minimise l'augmentation de l'inertie intra-classe (Ward).

24. Quel sera le résultat de l'opération suivante ?

On coupe l'arbre pour faire trois parties, par exemple à la hauteur de 15 pour faire trois classes. Le résultat est :

```
cutree(h0,3)
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
 1  1  1  1  1  1  1  1  1  1  2  2  2  2  2  2  2  2  2  2  2  2  3  3  3  3
27 28 29 30
 3  3  3  3
```

25. Donner quelques idées pour continuer l'analyse de ces données.

On pourrait interpréter la classification qu'on vient de faire des stations en trois classes en calculant les moyennes par classes des variables, reporter cette partition sur la carte de la rivière, s'en servir comme point de départ d'une classification en trois classes (kmeans) pour tester le solidité de cette simplification, refaire l'opération sur la carte de l'ACP du tableau des communautés de Poissons (fiche stage1, p.19), calculer les abondances moyennes par classes de stations de chacune des espèces, croiser avec une classification sur le tableau faunistique, calculer les corrélations entre coordonnées de l'ACP faune et de l'ACP milieu, …