SY09 Printemps 2009 - TP 1
Statistique descriptive, Analyse en composantes principales

Le principal but de ce TP est de nous découvrir les différentes utilisations des statistiques descriptives et de l'analyse en composant principale pour le traitement de tableaux de données. Ainsi l'objectif de la statistique descriptive est de décrire, c'est-à-dire de résumer ou représenter, par des statistiques, les données disponibles quand elles sont nombreuses.

1) Statistique descriptive

1.1 Données babies

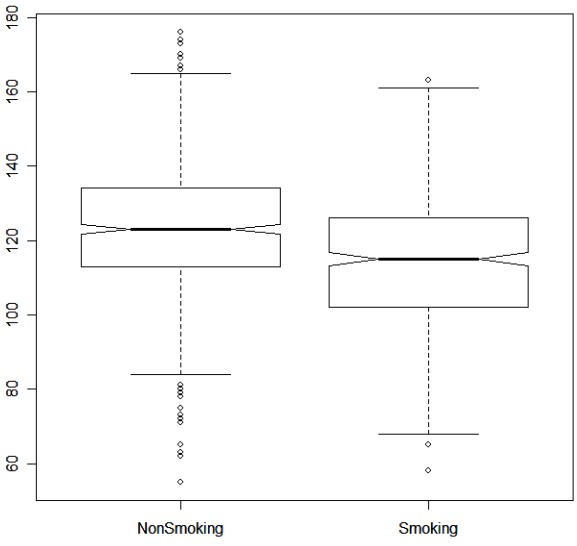
1. *Quelle est la différence de poids entre les bébés nés de mères qui fumaient durant leur grossesse et celles qui ne fumaient pas ?*

Nous comparons **une variable quantitative** (poids de naissance) et **une variable qualitative** (mère fumeuse durant la grossesse).

A partir de notre échantillon de 1236 naissances on a pu comparer le poids des bébés à la naissance en fonction du fait que la mère fumait ou pas.

En affichant les statistiques de bases pour les groupes Fumeur et NonFumeur, on se rend compte d'une légère différence au niveau du poids des bébés (variable bwt).En effet, la moyenne des poids des bébés NonSomking (c'est à dire dont les mères ne fume pas) est bien supérieur à celle des bébés Smoking. La moyenne des bébés NonSmoking est de 123.047 (on considérera ce poids comme étant le poids normale des bébés à la naissance) alors que celle des bébés Smoking est de 114.109

Mais cette différence est-elle significative ?



Deux modes de représentation permettent de répondre à cette question :

- Une représentation sous forme d'histogramme : On peut constater une différence (qui apparait aussi en faisant les moyennes) cependant, cela ne nous permet pas de définir une différence significative; puisque les histogrammes sont assez proches.
- Une représentation sous forme de « boite à moustaches » : en y ajoutant l'option « notch » qui affiche les intervalles de

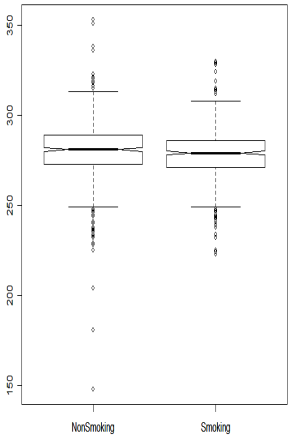
confiance. Cette représentation permet d'affirmer qu'il y a bien une différence significative de poids pour les bébés naissant d'une mère fumeuse et les bébés naissant d'une mère non-fumeuse car les intervalles de confiance ne se chevauchent pas.

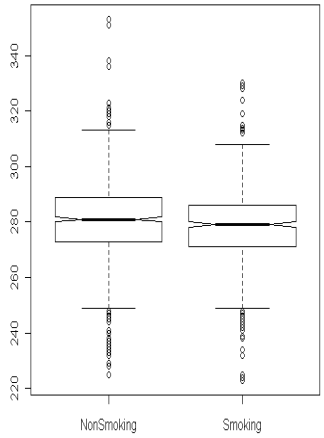
2. *Est-ce qu'une mère qui fume durant sa grossesse est encline à avoir un temps de gestation plus court qu'une mère qui ne fume pas ?*

Voici les moyennes des durées de gestation chez:

Mère Smoking : 277.9792 jours de gestation. Mère NonSmoking : 280.1869 jours de gestation.

Les moyennes des durées de gestation chez les mères Smoking et chez les mères NonSmoking sont très semblables. Visiblement, les mères Smoking et nonSmoking auraient une durée de gestation similaire. Pour vérifier ceci, nous nous baserons également sur les boites à moustaches en y ajoutant les notches. Ainsi nous observerons la similarité des médianes.





La représentation par boite à moustache révèle que la différence n'est pas significative car les intervalles de confiance se chevauchent.

Afin de pouvoir apporter une réponse à la question, nous allons faire maintenant une analyse statistique en déterminant l'intervalle de confiance sur la médiane pour les deux variables smoking et non smoking : Intervalle de confiance à 0,95

Intervalle de confiance	95%		90%	
	Smoking	NonSmoking	Smoking	NonSmoking
Moyenne :	277,98	280,19	277,97	280,18
Borne sup IC	276,63	278,98	276,84	279,18
Borne inf IC	279,33	281,39	279,11	281,18

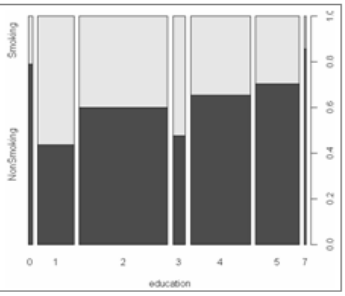
Nous pouvons donc affirmer (mais seulement à 90%) qu'une mère qui fume durant la grossesse encline à avoir un temps de gestation plus court qu'une mère qui ne fume pas.

3. *Le niveau d'étude a-t-il une influence sur le fait que la mère soit fumeuse ?*

Pour voir si le niveau d'étude à une influence sur le fait que la mère soit fumeuse, on a décidé de mettre en place un tableau de contingence pour évaluer pour chaque niveau d'étude le pourcentage de mère fumeuse.

Voici le tableau de contingence et sa représentation en histogramme :

```
> table(smoke,education)
          education
smoke       0   1   2   3   4   5   7   9
  NonSmoking 15  79 264  30 194 154   6   0
  Smoking     4 102 176  33 102  65   1   1
```



On a tout d'abord pensé à utiliser un test du chi² qui nous a informés sur le fait que les deux variables sont effectivement dépendantes. Cependant ce Test ne nous dit pas qu'elle variable influence l'autre.

Nous allons, donc, nous contenter d'une interprétation générale des chiffres de notre tableau de contingence grâce à des pourcentages pour chaque niveau d'étude.

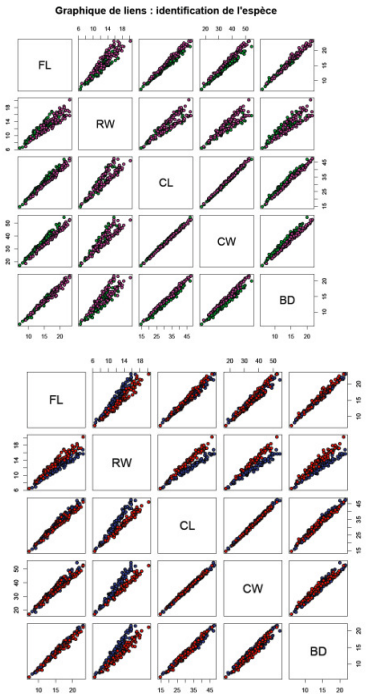
éducation	0	1	2	3	4	5	7	9
%NonSmoking	0,78	0,43	0,6	0,48	0,66	0,71	0,86	0
%Smoking	0,22	0,57	0,4	0,52	0,34	0,29	0,14	1

On peut remarquer que la taille des l'échantillon des niveaux d'étude 7 et 9 sont faible et que, donc, les pourcentages ne seront pas très représentatifs de la situation réelle. Cependant pour les autres niveaux d'étude la taille de l'échantillon est assez grande pour donner des conclusions. On remarque grâce au tableau précédent que plus le niveau d'étude est haut et plus le pourcentage de NonSmoking augmente (voir aussi l'histogramme, plus simple à remarquer) sauf pour les niveaux 1 et 3 d'éducation. On peut donc supposer que le niveau d'étude à une influence sur le fait que la mère soit fumeuse. On observe une différence significative après 4 ans d'études. Ainsi un niveau d'étude élevé réduit les chances que la mère soit fumeuse.

1.2 Données crabs

1) *Effectuer dans un premier temps une analyse descriptive des données. Existe-t-il des différences de caractéristiques morphologiques selon l'espèce ou le sexe ? Semble-t-il possible d'identifier l'espèce ou le sexe d'un crabe à partir d'une ou plusieurs mesures caractéristique ?*

Pour répondre la première question nous allons tracer les graphiques de corrélation.



On peut remarquer que les variables sont à peu près toutes corrélées. Donc, on ne peut pas affirmer existence d'une différence de caractéristique morphologique selon l'espèce ou le sexe.

Pour pouvoir pousser notre analyse plus loin nous avons effectué une analyse descriptive en fonction du sexe ou de l'espèce du crab. La méthode utilisée est la méthode des boxplots où nous représentions a chaque fois une variable quantitative (FL, RW….) en fonction d'une variable qualitative (sp,sex) avec l'intervalle de confiance.

Les résultats nous on permit de conclure qu'il y a bien une différence de caractéristiques morphologiques significative selon l'espèce du crab.

Cependant, selon le sexe, cette différence n'est significative que pour la caractéristique RW contrairement à l'espèce.

On conclut qu'il est possible d'identifier l'espèce du crab à partir d'une ou plusieurs mesures ce qui n'est pas le cas pour le sexe, puisque nous avons seulement une différence significative (RW) qui ne nous suffit pas pour connaître le sexe d'un crabs dans certain cas.

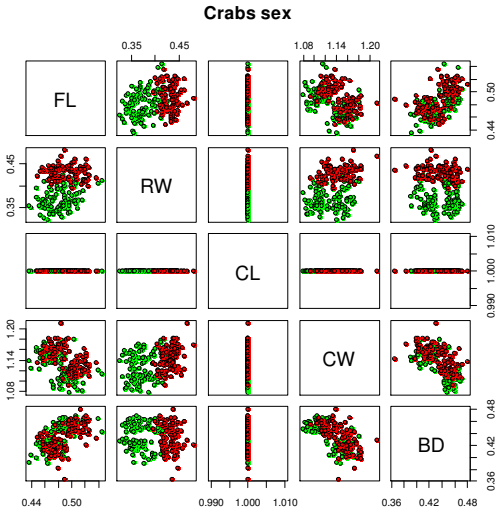
2) *Dans un second temps, on étudiera la corrélation entre les différentes variables. Quelle en est vraisemblablement la cause ? Quel traitement est-il possible d'appliquer aux données pour s'affranchir de ce phénomène ?*

La raison pour la quelle les variables sont à peu près toutes corrélées est qu'elles sont toutes plus au moins proportionnelle à la taille de l'animal, donc toute les caractéristiques sont liées.

La solution est de diviser par une des variables afin de la supprimer. Notre choix c'est posé sur la variable CL puisque c'est l'une des plus corrélée.

Ainsi pour connaître l'espèce de l'un des spécimens, il nous suffit d'avoir les combinaisons de variable suivantes : FL & CL, FL & CW ; RW & CW et RW & CL.

Ainsi pour connaître le sexe de l'un des spécimens, il nous suffit d'avoir les combinaisons de variable suivantes : FL & CL, FL & CW ; FL & BD, CL & CW, CL & BD et CW & CL.



2) Analyse en composantes principales

L'Analyse en Composantes Principales (ACP) est une méthode de la statistique multi variée, qui consiste à transformer des variables liées entre elles (dites "corrélées" en statistique) en nouvelles variables indépendantes les unes des autres (donc "non corrélées"). Ces nouvelles variables sont nommées "composantes principales", ou axes. Elle permet au praticien de réduire l'information en un nombre de composantes plus limité que le nombre initial de variables.

2.1 Exercice théorique

2.2 Traitement des données Crabs

1. *Calculer les axes factoriels de l'ACP du nuage ainsi défini. Quels sont les pourcentages d'inertie expliquée par chacun de ces axes.*

Nous appliquons la méthode d'analyse en composantes principales.
Nous centrons la matrice A en colonne. Nous calculons ensuite la matrice de variance S.

$$A=\begin{pmatrix}3 & 4 & 3\\ 1 & 4 & 3\\ 2 & 3 & 6\\ 2 & 1 & 4\end{pmatrix}$$ Matrice A centrée : $$Ac=\begin{pmatrix}1 & 1 & -1\\ -1 & 1 & -1\\ 0 & 0 & 2\\ 0 & -2 & 0\end{pmatrix}$$ Matrice de variance: $$S=\begin{pmatrix}0{,}5 & 0 & 0\\ 0 & 1{,}5 & -0{,}5\\ 0 & -0{,}5 & 1{,}5\end{pmatrix}$$

Nous pouvons alors calculer les matrices des **valeurs propres L et des vecteurs propres U.** On a:

$$L=\begin{pmatrix}2 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 0.5\end{pmatrix}$$ $$U=\begin{pmatrix}0 & 0 & 1\\ -0{,}70710 & 0{,}70710 & 0\\ 0{,}70710 & 0{,}70710 & 0\end{pmatrix}$$

Calcul des pourcentages d'inertie expliquée par chacun de ces axes :

Axe 1 : $\frac{2}{2+1+0.5}$ x 100 = 57.14286 % Axe 2 : $\frac{1}{2+1+0.5}$ x 100 = 28.57143 % Axe 3 : $\frac{0.5}{2+1+0.5}$ x 100 = 14.28571%

On remarque que l'axe 3 est moins intéressant que les deux premiers, puisqu'il explique seulement 14% de l'inertie générale.

2. *Calculer les composantes principales ; en déduire la représentation des quatre individus dans le premier plan factoriel.*

On obtient la matrice des composantes principales en utilisant la formule : C= Ac*U.

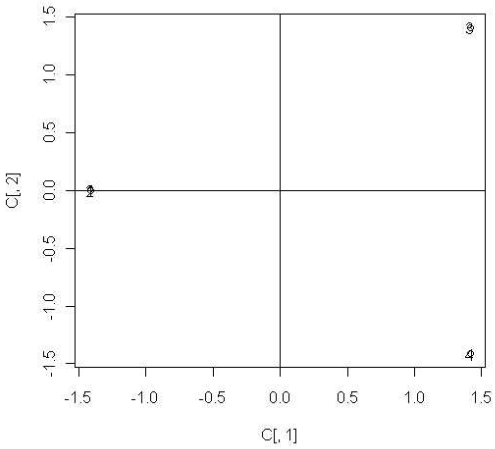
$$C = \begin{pmatrix} -1{,}414214 & 2.220446e-16 & 1 \\ -1{,}414214 & 2.220446e-16 & -1 \\ 1{,}414214 & 1{,}414214 & 0 \\ 1{,}414214 & -1{,}414214 & 0 \end{pmatrix}$$

3. *Tracer la représentation des trois variables dans le premier plan factoriel.*

Le premier plan factoriel est Δu1, composé des u1 et u2 de la matrice C.

les individus 1 et 2 sont confondus (que se soit vis-à-vis de C[,1] ou C[,2]), et les individus 3 et 4 sont similaires vis-à-vis de C[,1] mais vis-à-vis de C[,2].

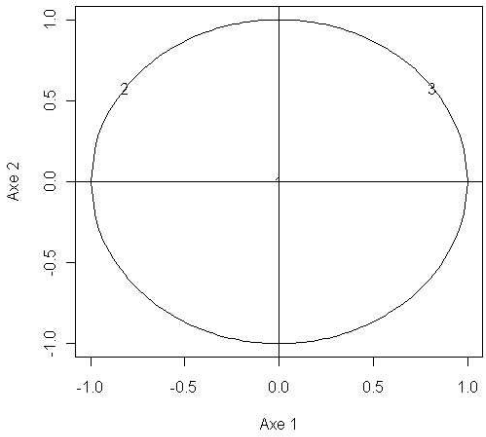
Le premier axe, C[,1], permet donc d'isoler les 2 premiers individus des 2 derniers. Le second, C[,2] permet permets ensuite de différencier l'individu 3 du 4.



Représentation des trois variables dans le premier plan factoriel

Les variables 2 et 3 sont similaires sur l'axe 2 et qu'elles sont bien représentées (proche du cercle).

La variable 1 est mal représentée, elle n'est ni corrélée par l'axe 1, ni corrélée par l'axe 2 ; elle sera alors sûrement expliquée par un autre plan factoriel.



4. *Calcul de l'expression $\sum_{\alpha=0}^{k} c\ \sigma\ u'\sigma$ pour k = 1, 2, 3*

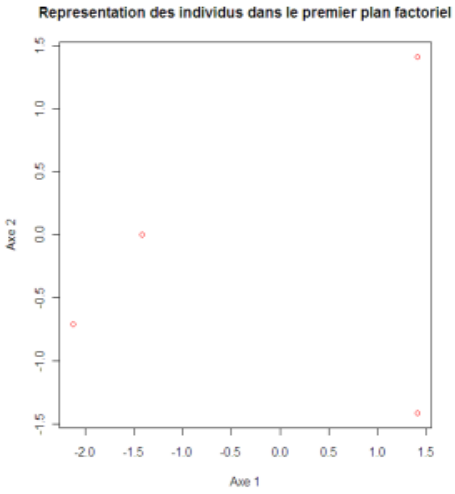
k = 1 : $c\sigma\ u'\sigma : \begin{pmatrix} 0 & 1 & -1 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \\ 0 & -1 & 1 \end{pmatrix}$

k = 2 : $\sum_{\alpha=0}^{2} c\ \sigma\ u'\sigma : \begin{pmatrix} 0 & 1 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 2 \\ 0 & -2 & 0 \end{pmatrix}$ k = 3 : $\sum_{\alpha=0}^{3} c\ \sigma\ u'\sigma :$

$\begin{pmatrix} 1 & 1 & -1 \\ -1 & 1 & -1 \\ 0 & 0 & 2 \\ 0 & -2 & 0 \end{pmatrix}$ = Matrice A centrée

5. *Représenter dans leur plan factoriel respectif l'individu supplémentaire (4, 4, 2) et la variable supplémentaire (6, 7, 3, 4).*

Pour représente ce nouveau individu il faut soustrait la moyenne de chaque colonne de la matrice initiale Y à ses coordonnées. Ensuite, on projette ses coordonnées sur les trois axes factoriels. Ce qui nous donne le vecteur (-2.121, - 0.7071).



1.2 Traitement des données Crabs

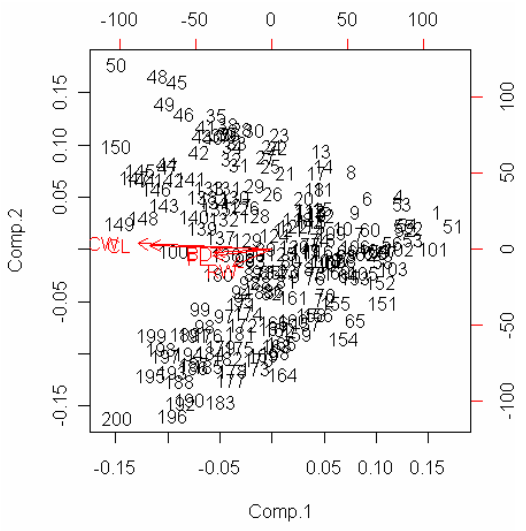
A l'aide de l'ACP sans traitement préalable nous allons essayer de trouver une représentation des données qui permette de distinguer les crabes selon leur espèce et leur sexe

1. *Tester tout d'abord l'ACP sur crabs quant sans traitement préalable. Que constatez-vous ?*

Cette fois ci, nous avons travaillé directement avec l'ACP par le code R princomp.

On obtient la représentation graphique de ces données par le code R : biplot(xx) qui représente dans un même graphique les 200 individus et les 5 variables.

Avec ce type de graphique, nous ne pouvons pas distinguer visuellement différents groupes de crabs lié à l'espèce et au sexe. Nous ne pouvons trouver que tous les variables sont bien orientés ver la première composant. Ce qui montre que toute l'information est représentée sur la première l'axe.

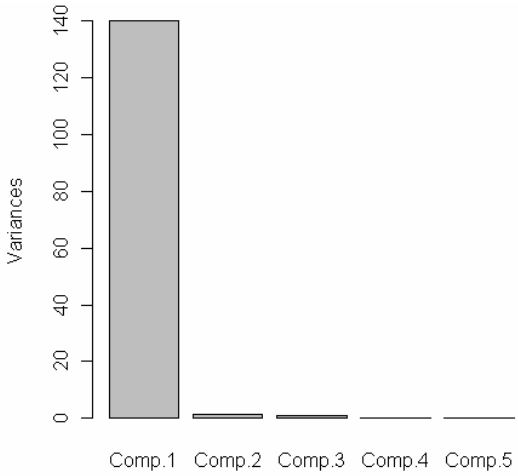


2. *Trouver une solution pour améliorer la qualité de votre représentation en termes de visualisation des différents groupes.*

D'après la première question tous les 5 variables sont orientés vers le premier composant.

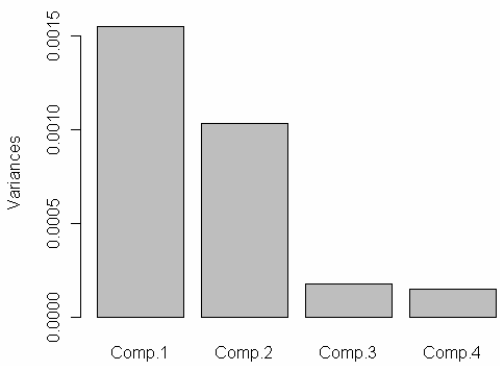
Ce qui nous permet de prétendre qu'il y a une forte corrélation sur le premier axe dû à l'effet de taille.

Tous ces variables sont liés par certaines informations. Ce qui est normal s'il y a des corrélations entre chacun des variables. La mission maintenant est donc de limiter le variable dont les valeurs propres sont les plus élevées afin de donner la même importance à chaque variable et à chaque individu. Voici le tableau de corrélation : cor(crabsquant)

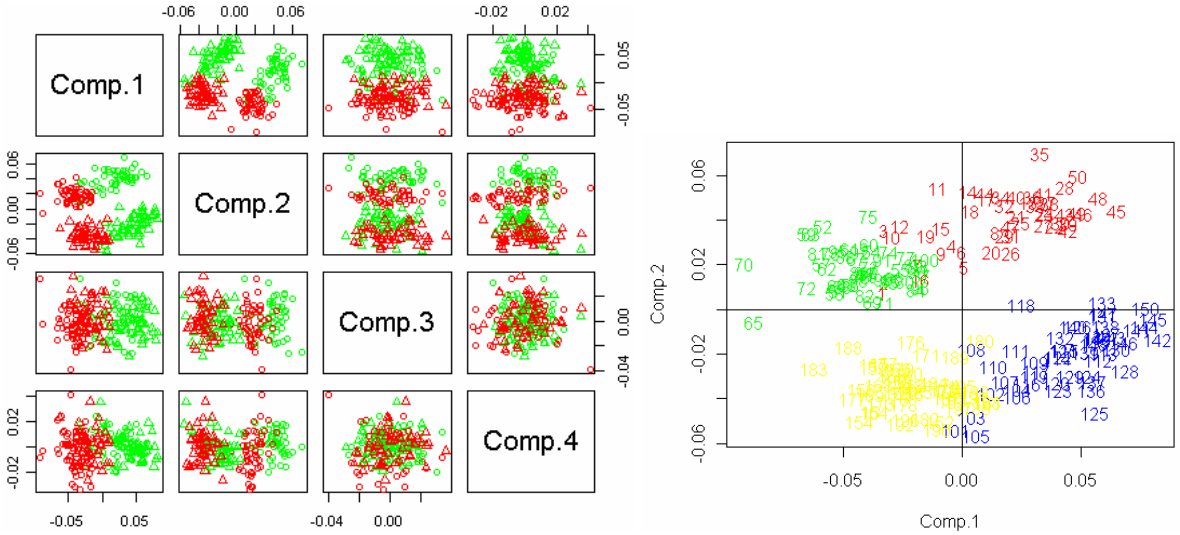


	FL	RW	CL	CW	BD
FL	1.0000000	0.9069876	0.9788418	0.9649558	0.9876272
RW	0.9069876	1.0000000	0.8927430	0.9004021	0.8892054
CL	0.9788418	0.8927430	1.0000000	0.9950225	0.9832038
CW	0.9649558	0.9004021	0.9950225	1.0000000	0.9678117
BD	0.9876272	0.8892054	0.9832038	0.9678117	1.0000000
Σ	4,8384124	4,5893381	**4,8498111**	4,8281921	4,8278481

Voilà notre résultat : Nous avons plus de dominant d'un variable : plot(xx)



Dans le graphique suivant, dès le premier plan, nous avons déjà distingué les 4 groupes de crabes par espèce et par sexe (BM – BF – OM – OF)



3) Conclusion

Nous avons apprit au début qu'il est assez compliqué de définir la meilleure description possible d'un phénomène. Les méthodes exploratoires avec l'utilisation des différentes représentations graphiques (Boxplot, histogramme,…) comportent toutes des avantages et des inconvénients.

Puis sur deuxième partie nous avons appris à utiliser la méthode d'analyse en composantes principales, et ce par calcul matriciel et en mode automatique avec la fonction princomp du logiciel R.

Finalement nous avons conclue qu'il faut avoir un esprit critique pour garantir des résultats concluant lors de nos analyses.