SY09 - TP01
Statistique descriptive, Analyse en composantes principales

Bertrand Bon - Antoine Hars

April 22, 2015

L'objectif de ce tp est d'apprendre à manipuler une grande quantité de données au moyen d'outils de statistique descriptive et d'analyse en composantes principales.

1 Statistique Descriptive

1.1 Données babies

Dans cet exercice, nous disposons d'un jeu de données *babies* constitué de 1236 bébés décrits par 23 variables. Dans notre tp, nous n'utilisons que 8 variables : 5 quantitatives (le poids à la naissance, la durée de gestation, le nombre de grossesses précédentes, la taille de la mère et le poids de la mère) et 3 qualitatives (l'âge de la mère, si la mère fume ou non et le niveau d'éducation de la mère).

Quelle est la différence de poids entre les bébés nés de mères qui fumaient durant leur grossesse et celles qui ne fumaient pas ?
Pour observer cette différence, nous pouvons étudier le résumé des valeurs du poids des bébés en fonction du fait que leur mère fumait ou non :

Min	1st Qu.	Median	Mean	3rd Qu.	Max
58.0	102.0	115.0	114.1	126.0	163.0

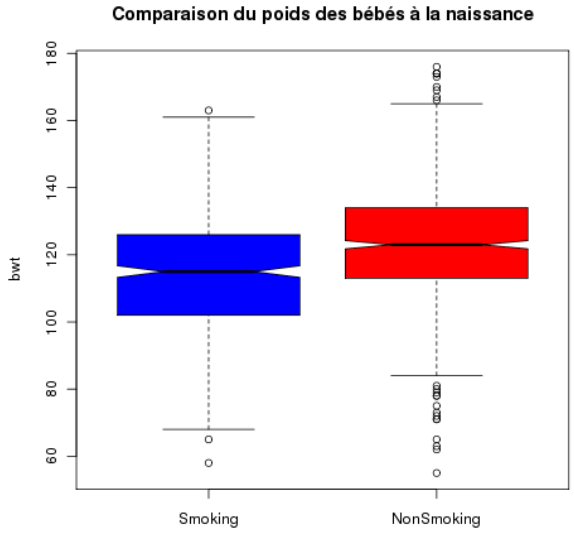
Résumé des valeurs des bébés nés d'une mère fumeuse

Min	1st Qu.	Median	Mean	3rd Qu.	Max
55.0	113.0	123.0	123.0	134.0	176.0

Résumé des valeurs des bébés nés d'une mère non fumeuse

À travers ces deux tableaux, nous pouvons observer une différence entre les deux populations : Suivant la moyenne et la médiane des deux tableaux de données, il semble que les enfants de mère fumeuse ont tendance à naître plus maigres que ceux d'une mère non fumeuse.

Pour vérifier cette tendance, nous pouvons observer le graphique suivant :



L'observation de ce boxplot nous permet de confirmer la tendance observée. De plus, nous pouvons constater une présence plus importante de valeurs atypiques chez les mères fumeuses.
Les intervalles de confiance pour les deux médianes ne se chevauchent pas donc les deux médianes différent et nous pouvons dire que dans 95% des cas, la différence de poids est significative. Un bébé d'une mère fumeuse est donc statistiquement plus maigre qu'un bébé d'une mère non fumeuse.

Est-ce qu'une mère qui fume durant sa grossesse est encline à avoir un temps de gestation plus courts qu'une mère qui ne fume pas ?
Pour répondre à cette question, nous étudions le résumé du temps de gestation en fonction du fait que la mère fumait ou non :

Min	1st Qu.	Median	Mean	3rd Qu.	Max
223.0	271.0	279.0	278.0	286.0	330.0

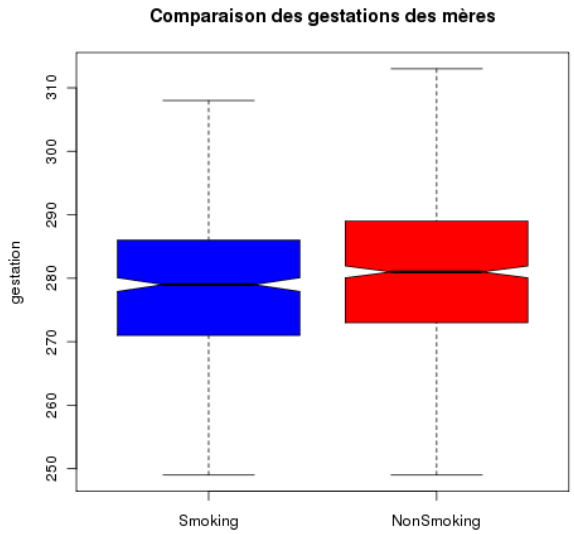
Résumé des valeurs du temps de gestation d'une mère fumeuse

Min	1st Qu.	Median	Mean	3rd Qu.	Max
148.0	273.0	281.0	280.2	289.0	353.0

Résumé des valeurs du temps de gestation d'une mère non fumeuse

Nous pouvons observer que les médianes sont légèrement différentes ce qui nous permet de dire que le fait de fumer peut conduire à une modification sur la gestation mais il n'est pas possible d'établir clairement ce point avec les deux résumés.

Nous avons ensuite observé les deux populations au moyen d'un boxplot :



Concernant ce graphique, il est apparu que du coté des femmes ne fumant pas, il y avait un nombre conséquent de valeurs atypiques. Afin d'étudier ce boxplot au niveau des médianes et des intervalles de confiance, nous n'affichons pas ces valeurs atypiques sur ce graphique.
Ce graphique nous montre que les intervalles de confiance des médianes se chevauchent, donc il n'est pas possible depuis ce graphique de dire si oui ou non les mères ne fumant pas ont un temps de gestation plus long que des mères fumeuses.

Le niveau d'étude a-t-il une influence sur le fait que la mère soit fumeuse ?
Afin de répondre à cette question, nous étudions dans un premier temps le tableau de contingence des données :

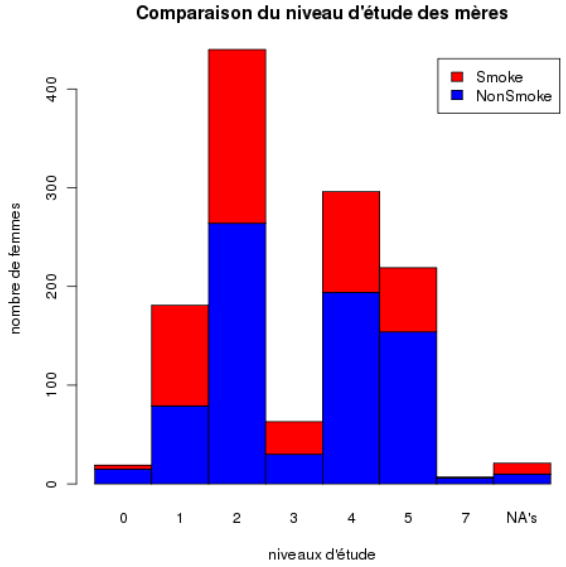
	0	1	2	3	4	5	7
Mères non fumeuses	15	79	264	30	194	154	6
Mères fumeuses	4	102	176	33	102	65	1

Tableau de contingence du niveau d'étude des mères

L'étude de tableau de contingence nous apporte des éclaircissements sur plusieurs points :
Nous pouvons affirmer qu'il y a de grande différences pour chaque classe d'étude si un mère fume ou pas.
Ce tableau ne prend pas en compte le niveau d'étude n°6 puisqu'aucune mère des deux populations n'en fait partie.

Concernant les niveaux d'étude 2, 4 et 5 (les niveaux avec le plus grand écart de valeurs), le nombre de mères varie beaucoup suivant le fait qu'elles fument ou non.

Nous pouvons ensuite observer les deux populations de mères au moyen d'un barplot :

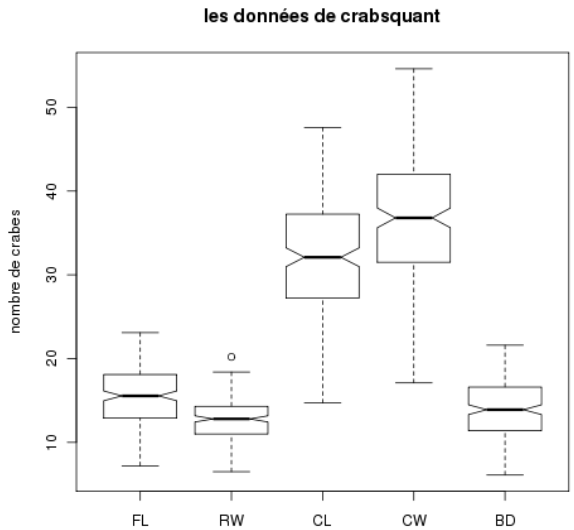


Ce barplot illustre le tableau de contingence et montre que pour les niveaux d'étude 2, 4 et 5, il y a une plus grand proportions de mères non fumeuses alors quand le cas du niveau d'étude 1, le nombre de mères fumeuses et plus grand que celui des mères non fumeuses.

L'étude présentée nous montre que les femmes non fumeuses sont moins propices à avoir un enfant avec un faible poids à la naissance que les femmes qui fument, ce qui correspond à nos analyses.
Du coté de l'étude présentée, il est expliqué qu'il n'y a pas de lien entre le fait de fumer et le temps de gestation des femmes. De notre coté, nous avons pu observer une légère différence entre les données des mères fumeuses et non fumeuses mais ces données ne nous permettent aucunement de déterminer statistiquement l'influence du tabagisme sur ce temps de gestation.
Concernant la relation entre le niveau d'étude et le tabagisme, nous pouvons dire qu'il semble que les femmes avec un niveau d'étude élevé fument moins mais il est apparu que sur notre échantillon de données, les femmes avec un niveau d'étude 2 et non fumeuses sont plus importantes que les femmes du même niveau et fumeuses donc il est difficile de trouver un lien entre ces deux critères.

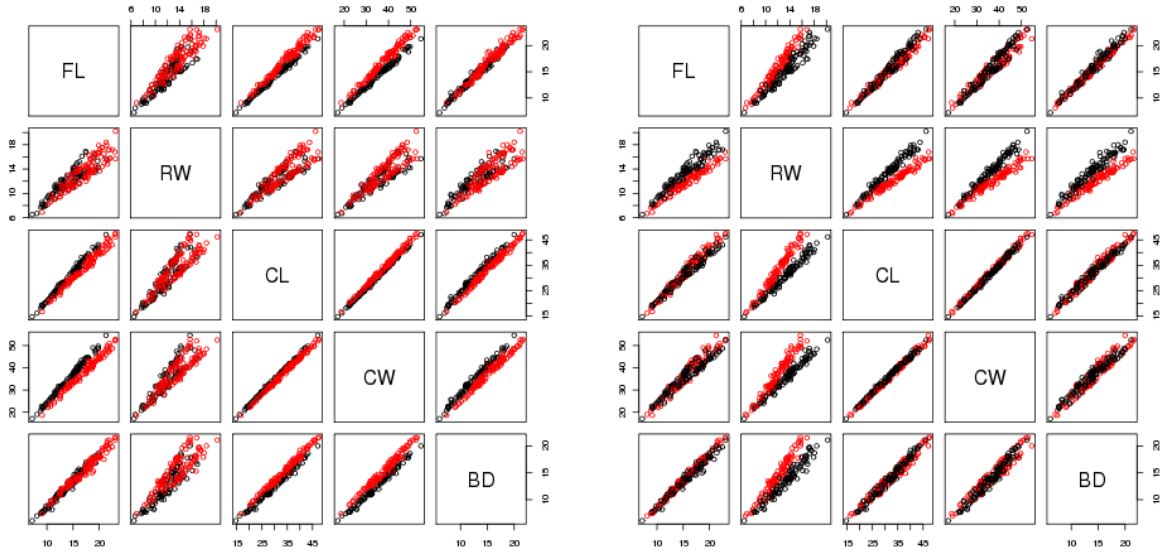
1.2 Données crabs

Dans cet exercice, nous avons étudié un jeu de données constitué de 200 crabes décrits par huits variables (trois variables qualitatives et cinq quantitatives).
Nous pouvons représenter les variables quantitaives de crabsquant sous la forme du graphique suivant :

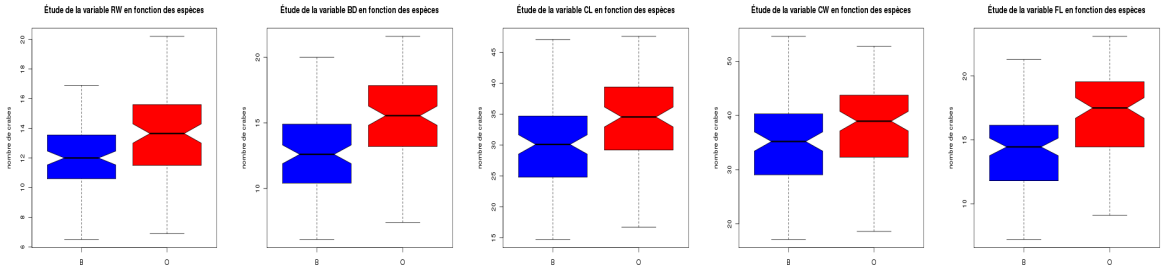


Ce graphique nous montre donc un valeur atypique pour la variable RW et fait ressortir la présence de deux groupes de variables ayant une dispersion similaire entre elles : les variables FL, RW et BD d'un coté et les variables CL et CW de l'autre.

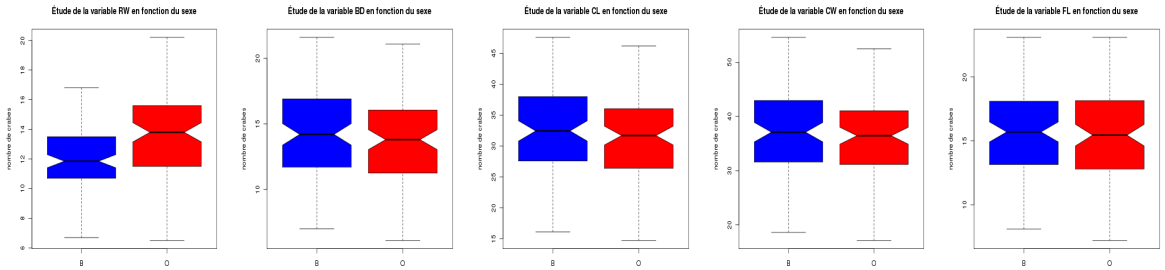
Existe-t-il des différences de caractéristiques morphologiques selon l'espèce ou le sexe ? Semble-t-il possible d'identifier l'espèce ou le sexe d'un crabe à partir d'une ou plusieurs mesures de ces caractéristiques ?
Nous affichons les données des crabes en fonction de l'espèce et du sexe des crabes :



Ces deux graphiques nous montrent qu'il est difficile de déterminer le sexe ou l'espèce d'un crabe à partir des variables disponibles.
Afin de déterminer l'existence de différences de caractéristiques morphologiques selon l'espèce ou le sexe, nous comparons les données suivantes :



En fonction des espèces, nous pouvons voir que toutes les caractéristiques diffèrent. La dispersion des données est par contre similaire pour chaque boxplot.



Par rapport au sexe des crabes, nous remarquons que toutes les caractéristiques (à part RW) sont plutôt proches au vu du chevauchement des intervalles de confiance et la dispersion des données est environ la même pour les boxplots.

Les crabes B ont des valeurs plus élevées pour toutes les variables quantitatives.

Quelle la cause de la corrélation entre les différentes variables ? Quel traitement est-il possible d'appliquer aux données pour s'affranchir de ce phénomène de corrélation ?
On peut voir qu'il existe une forte corrélation entre toutes les combinaisons de variables. Vu qu'il s'agit de mesures de certaines parties du corps des crabes, il est normal que chaque variable soit proportionnelle aux autres. Le contraire signifierait que les crabes n'ont pas de proportions corporelles harmonieuses.

Pour confirmer cette forte corrélation, nous calculons les coefficients de corrélation :

	FL	RW	CL	CW	BD
FL	1.000	0.907	0.979	0.965	0.988
RW	0.907	1.000	0.893	0.900	0.889
CL	0.979	0.893	1.000	0.995	0.983
CW	0.965	0.900	0.995	1.000	0.968
BD	0.988	0.889	0.983	0.968	1.000

Tableau des coefficients de corrélation entre les variables

La variable la plus corrélée est la variable *CL* suite à la somme de ces corrélations. Pour s'affranchir de ce phénomène de corrélation, nous pouvons diviser les données par cette variable pour décorréler les variables.

2 Analyse en composantes principales

2.1 Exercice théorique

Le but de cette partie est de comprendre l'ACP, une analyse permettant de traiter des données multi-dimensionnelles d'un espace large de variables en réduisant celui-ci.

$$M = \begin{pmatrix} 3 & 4 & 3 \\ 1 & 4 & 3 \\ 2 & 3 & 6 \\ 2 & 1 & 4 \end{pmatrix}$$

Calcul des axes factoriels de l'ACP du nuage défini
Pour obtenir les axes factoriels, on centre la matrice :

$$M' = \begin{pmatrix} 1 & 1 & -1 \\ -1 & 1 & -1 \\ 0 & 0 & 2 \\ 0 & -2 & 0 \end{pmatrix}$$

puis on calcule la matrice de variance :

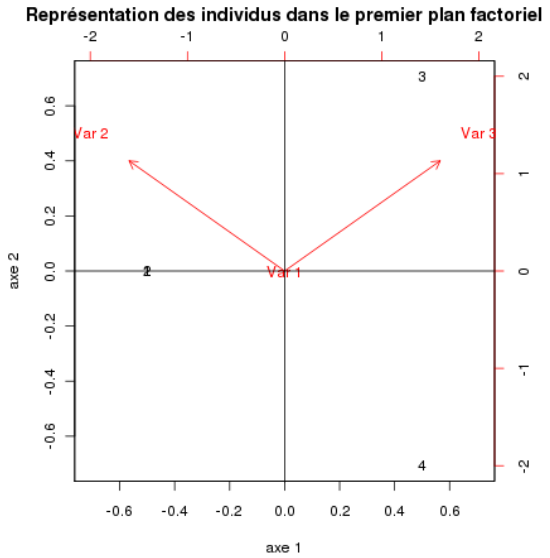
$$S = \frac{1}{n}.M'.M = \frac{1}{4}.M'.M = \begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 1.5 & -0.5 \\ 0 & -0.5 & 1.5 \end{pmatrix}$$

En diagonalisant cette matrice, nous obtenons les valeurs propres et les axes d'inertie suivants :

	$\lambda 1$	$\lambda 2$	$\lambda 3$
valeurs propres	2.0	1.0	0.5
% axes d'inertie	57.14	28.57	14.29
% axes d'inertie cumulés	57.14	85.71	100.0

Nous pouvons remarquer que les deux premiers axes cumulent environ 86% de l'information. Donc nous pouvons représenter 86% de l'information sur le plan factoriel défini par les deux premiers axes.
Le calcul des composantes principales donne la matrice :

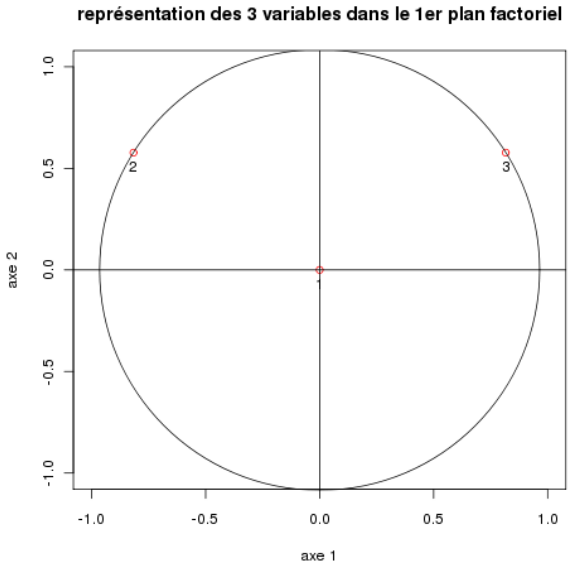
$$C = \begin{pmatrix} -1.41 & 0 & 1 \\ -1.41 & 0 & -1 \\ 1.41 & 1.41 & 0 \\ 1.41 & -1.41 & 0 \end{pmatrix}$$



Sur ce graphique, nous pouvons observer que les deux premiers individus correspondent au même point dans le premier plan factoriel et que la seule coordonnée qui peut différencier ces deux points dépend du troisième axe factoriel.

Traçage de la représentation des trois variables dans le premier plan factoriel
On calcule les corrélations entre les variables pour avoir leurs coordonnées sur le premier plan factoriel :

$$D = cor(M', C) = \begin{pmatrix} 0 & 0 & 1 \\ -0.816 & 0.577 & 0 \\ -0.816 & 0.577 & 0 \end{pmatrix}$$

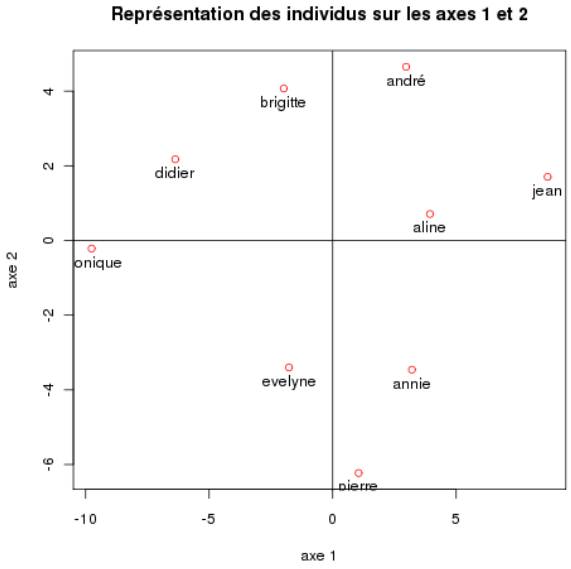


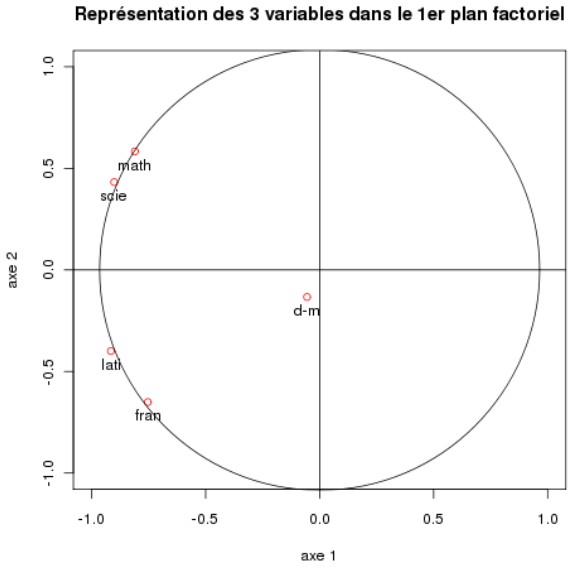
Calcul de l'expression $\sum_{\alpha=0}^{n} c_\alpha . u'_\alpha$ pour les valeurs $k = 1$, 2 et 3

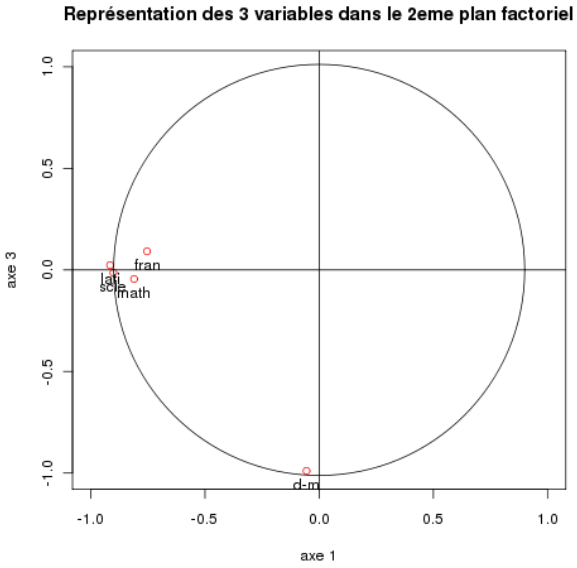
$$k = 1 : \begin{pmatrix} 0 & 1 & -1 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \\ 0 & -1 & 1 \end{pmatrix}, k = 2 : \begin{pmatrix} 0 & 1 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 2 \\ 0 & -2 & 0 \end{pmatrix}, k = 3 : \begin{pmatrix} 1 & 1 & -1 \\ -1 & 1 & -1 \\ 0 & 0 & 2 \\ 0 & -2 & 0 \end{pmatrix} = M'$$

2.2 Utilisation des outils R

Effectuer l'ACP du jeu de données notes étudié en cours. Montrer comment on peut retrouver tous les résultats alors obtenus (valeurs propres, axes principaux, composantes principales, représentations graphiques)
On calcule l'ACP avec l'instruction : *acp = princomp(notes)*
On obtient les axes d'inertie et axes d'inertie cumulés avec l'instruction *summary(acp)*
La matrice des composantes principales est contenue dans la variable *acp$scores*
Les vecteurs propres sont obtenue dans la variable *acp$loadings*
On établit le graphique de la représentation des individus :







Qu'affichent les fonctions *plot* et *biplot* ?
La fonction *princomp* réalise l'ACP sur la matrice et nous retourne l'écart-type pour la valeur *sdev*. La valeur *loadings* nous permet d'avoir les axes factoriels, c'est-à-dire les vecteurs propres de la matrice de variance. La valeur *scores* nous donne la matrice des composantes principales.

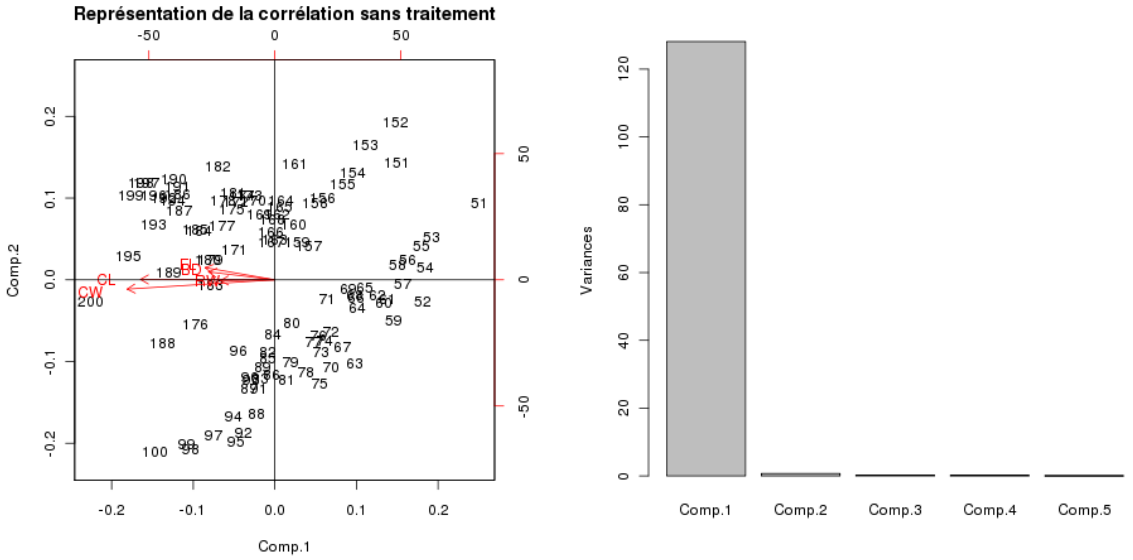
L'utilisation de la fonction *plot* sur le résultat de *princomp* affiche les valeurs propres associées à chaque composante.
La fonction *biplot* permet de projeter les individus et les variables sur un même plan. Il est utile d'utiliser cette fonction pour évaluer graphiquement les corrélations entre les variables (par rapport à l'angle que forment deux vecteurs). Deux variables sont indépendantes si leur vecteur forment un angle de 90°.

La fonction *biplot.princomp* donne accès à des options supplémentaires par rapport à la fonction *biplot*. En argument nous avons l'objet de la classe *princomp*, nous avons aussi la valeur *choices* pour définir la taille des vecteurs pour le plot. Nous avons aussi la valeur *scale* pour obtenir une représentation

standard des données. Pour finir, il y a la valeur *pc.biplot* qui si elle est mise à TRUE, réfère à un plot avec des observations élargies par la racine carrée des n et des variables réduites par cette racine carrée.

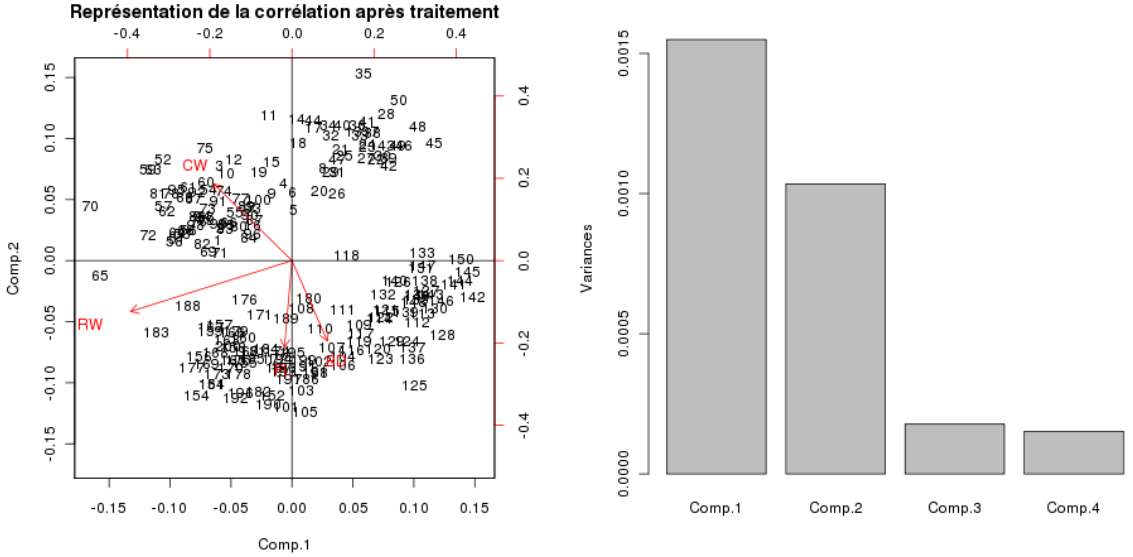
2.3 Traitement des données Crabs

Test de l'ACP sur *crabsquant* sans traitement préalable. Que constatez vous ?
La représentation de l'ACP sans traitement préalable nous donne le graphique suivant :

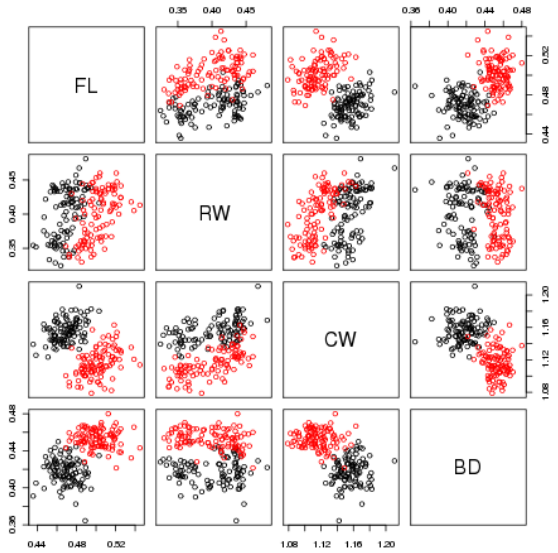


Nous pouvons observer sur ce graphique la forte corrélation des variables comme vu dans les questions précédentes.

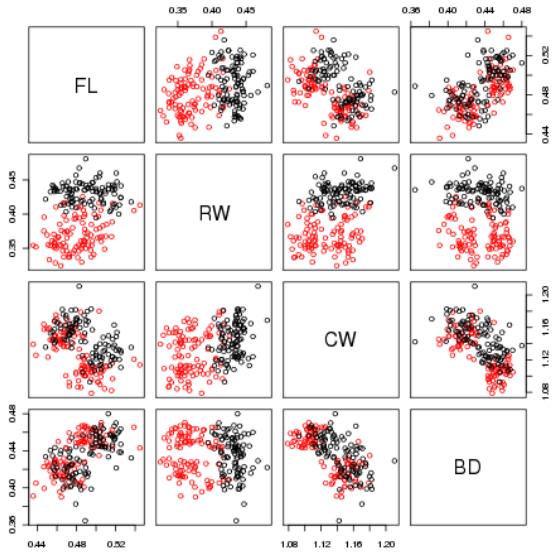
Trouver une solution pour améliorer la qualité de votre représentation en termes de visualisation des différents groupes
Pour améliorer la qualité de la représentation en termes de visualisation, nous avons retiré la variable *CL* des données puis nous avons refait une ACP et nous avons représenté les données obtenues :



Nous pouvons remarquer que les variables ne sont plus aussi corrélées que précédemment ce qui nous donne les graphiques suivants pour différencier les espèces :



Le graphique pour différencier les sexes des crabes est :



Conclusion :
Nous avons pu voir que l'ACP est un outil utile pour représenter des informations qui auraient été difficiles à détecter avec seulement l'utilisation de la statistique descriptive.