SY 09 - P14
TP 1 : Statistique descriptive, Analyse en composantes principales

Ricard Tatiana, Mehr Jean-Christophe

OBJECTIF DU TP

Les statistiques permettent depuis longtemps d'interpréter les données de manière objective et d'attribuer aux résultats un certain degré de confiance.
Les outils de mesure et de récolte de données actuels nous permettent d'engranger un nombre colossal d'information. Il est donc devenu nécessaire d'être capable de traiter ces données et d'y appliquer les outils de visualisation adéquats.
Lors de ce TP nous allons tenter d'extraire des informations spécifiques à partir de différents jeux de données. En utilisant des outils appropriés et en appliquant un traitement préalable aux données si nécessaire.

1 Statistique Descriptive

1.1 Données babies

Le jeu de données considéré **babies23.data** est constitué de 1236 bébés décrits par 23 variables. Une fois le jeu de données chargé, nous sélectionnons 8 variables:
- 5 quantitatives (le poids de naissance (birth weight), la durée de gestation, le nombre de grossesses précédentes, la taille de la mère et le poids de la mère)
- 3 qualitatives (l'âge de la mère, si la mère fume ou non et le niveau d'éducation de la mère).

Q1) 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 du caractère fumeur de la mère :

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

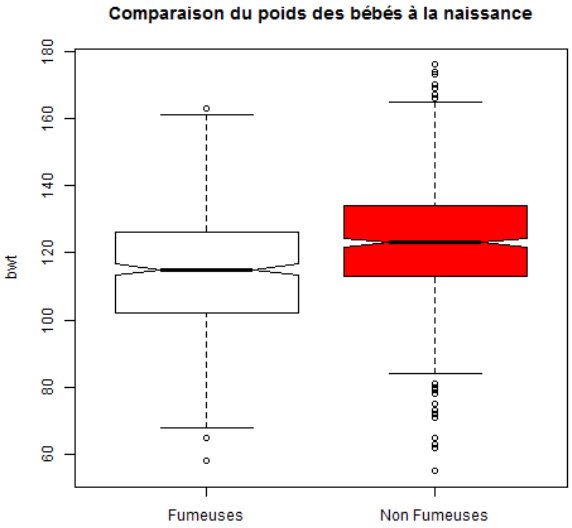
Résumé des données concernant une mère fumeuse

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

Résumé des données concernant une mère non fumeuse

Nous observons ainsi 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 aient un poids plus faible à la naissance 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, nous pouvons dire que la différence de poids est significative à 95%.

Q2) 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?

Pour répondre à cette question, nous étudions le résumé du temps de gestation en fonction du caractère fumeur de la mère :

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

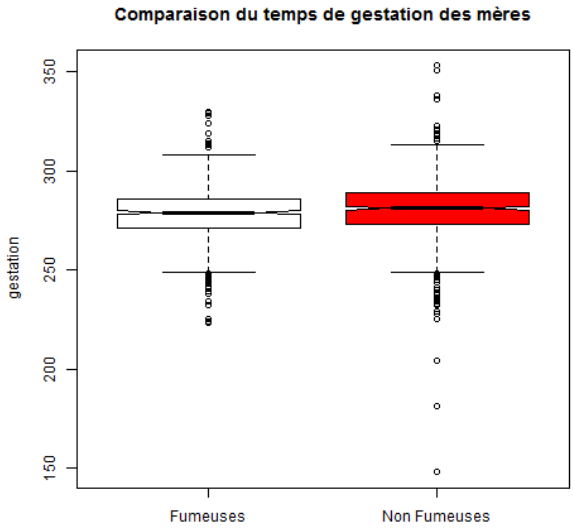
Résumé des données du temps de gestation d'une mère fumeuse

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

Résumé des données 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 :

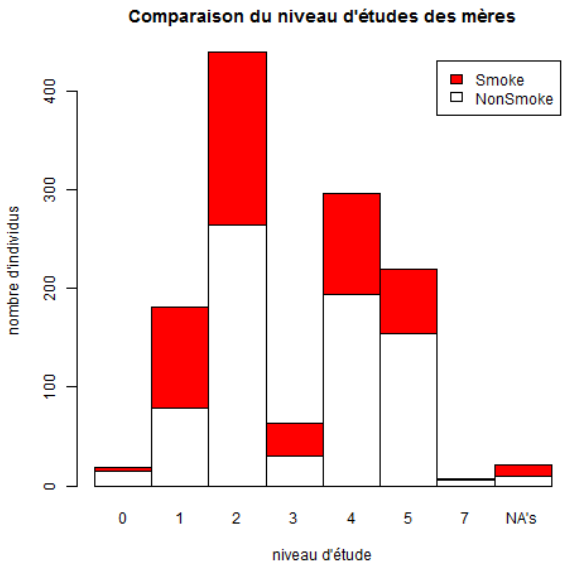


On constate ici un nombre important de valeurs atypiques chez les femmes non-fumeuses. Afin d'étudier ce boxplot au niveau des médianes et des intervalles de confiance, nous choisissons de ne pas afficher ces valeurs afin d'avoir une meilleur représentativité de l'échantillon. .
On observe ici un chevauchement des intervalles de confiance des médianest, il n'est donc pas possible de conclure à une significativité de la différence de temps de gestation entre les mères fumeuse et non-fumeuses.

Q3) 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

Le tableau de contingence nous permets de supposer un grand écart d'effectif entre le nombre de mère fumeuse et non fumeuse sur les niveaux d'étude 2, 4 et 5 (niveaux d'étude les plus représentés).
Nous choisissons ensuite d'observer les deux populations de mères au moyen d'un barplot :

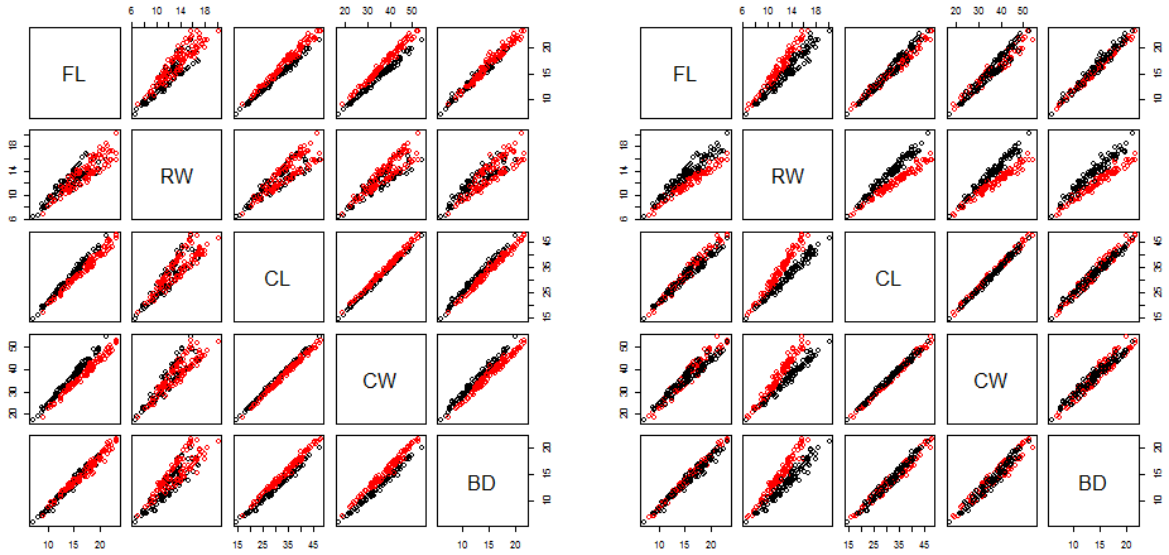


Ce barplot illustre le tableau de contingence et fait apparaître visuellement que pour les niveaux d'étude 2, 4 et 5, il y a une plus grand proportions de mères non fumeuses alors quand le cas du niveau d'étude 1, Cette tendance est inversée.
En comparant nos résultats à l'étude présentée,il est confirmé que les femmes non fumeuses ont tendance à avoir des bébés de poids plus élevé que les fumeuses.
L'étude évoque également l'absence de lien entre le temps de gestation et le caractère fumeur, ce qui explique que nous n'ayons pas pu déterminer de différence significative pour ce critère.
Concernant la relation entre le niveau d'étude et le tabagisme, nous remarquons à la rédaction de ce rapport qu'il aurait fallu effectuer la même étude en ramenant chacune des valeurs à un pourcentage. En effet, ici le nombre de femme non fumeuse évalué est de 742 alors que le nombre de femme fumeuse est inférieur à 500. Il est donc normal d'observer un nombre inférieur de femme fumeuse dans chaque catégorie d'étude. En prenant en compte ce fait, il semble tout de même que la proportion de femme fumeuse ayant un niveau d'étude supérieur à 4 est plus élevé.

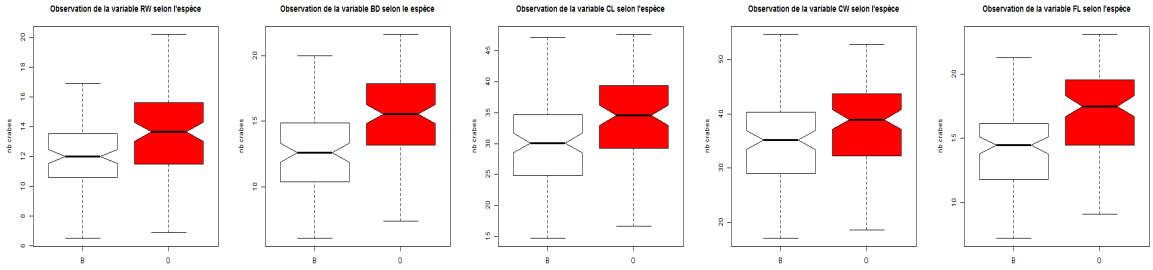
1.2 Données crabs

Le jeu de données considéré, disponible dans la bibliothèque de fonctions MASS, est constitué de 200 crabes décrits par huit variables (trois variables qualitatives, et cinq quantitatives).

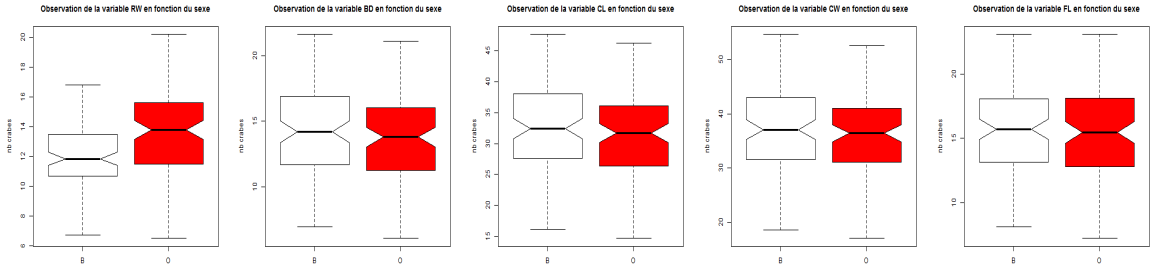
Q1) Existe-t-il des différences de caractéristiques morphologiques selon l'espèce ou le sexe? Semble-t-il possible d'identifier l'espèce ou le sexe d'un crabe à partir d'une ou plusieurs mesures de ces caractéristiques?
Nous affichons les données des crabes en fonction de l'espèce et du sexe des crabes :



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.

Q2) Quelle est vraisemblablement la cause de corrélation entre les variables ? Quel traitement est-il possible d'appliquer aux données pour s'affranchir de ce phénomène? Quel traitement est-il possible d'appliquer aux données pour s'affranchir de ce phénomène de corrélation ?
Nous calculons dans un premier temps les coefficients de corrélation via la commande **cor** :

	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

On a pu observer avec les différents outils appliqués une forte corrélation entre toutes les variables envisagées. Ce phénomène est dû à la nature des variables. En effet ce sont des variable morphologique liées à la taille du crabe.

On observe que la variable la plus corrélée est la variable CL suite à la somme de ces corrélations. Pour s'affranchir de ce phénomène de corrélation, il suffit de diviser les données par cette variable, les variables restantes devraient ainsi être affranchies de l'influence de la taille générale de l'individu.

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 :

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

puis on calcule la matrice de variance :

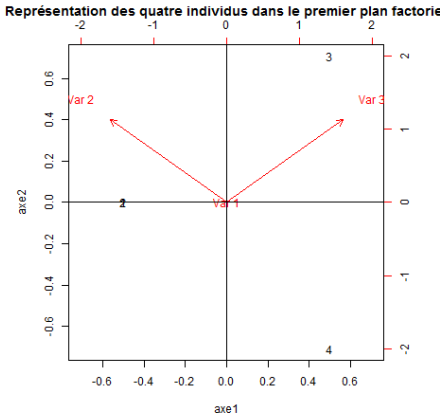
$$S = \frac{1}{n} * Mc * M = \frac{1}{4} * Mc * M = \begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 1.5 & -0.5 \\ 0 & -0.5 & 1.5 \end{pmatrix}$$

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 :

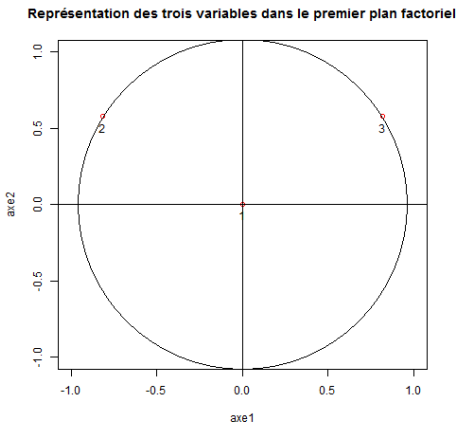
$$C = \begin{pmatrix} -1.41 & 0 & 1 \\ -1.41 & 0 & -1 \\ 1.41 & 1.41 & 0 \\ 1.41 & -1.41 & 0 \end{pmatrix}$$



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 :

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



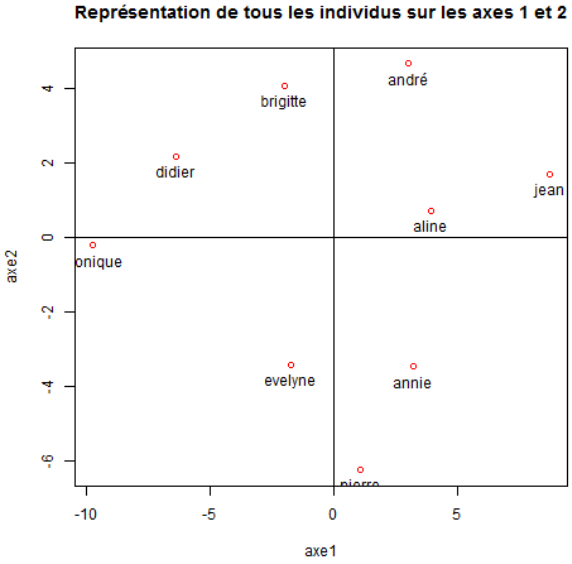
Calcul de l'expression $\sum_{\alpha=0}^{n} c_\alpha . u'_\alpha$ pour les valeurs $k = 1$, 2 et 3

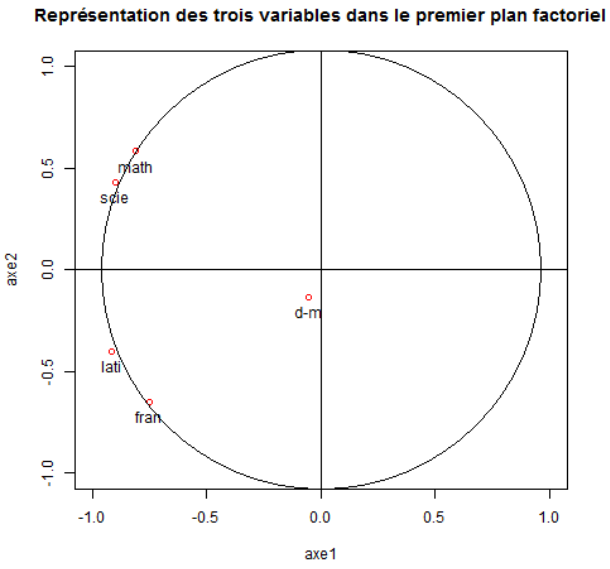
$$k = 1 : \begin{pmatrix} 0 & 1 & -1 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \\ 0 & -1 & 1 \end{pmatrix}, \; k = 2 : \begin{pmatrix} 0 & 1 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 2 \\ 0 & -2 & 0 \end{pmatrix}, \; k = 3 : \begin{pmatrix} 1 & 1 & -1 \\ -1 & 1 & -1 \\ 0 & 0 & 2 \\ 0 & -2 & 0 \end{pmatrix} = Mc$$

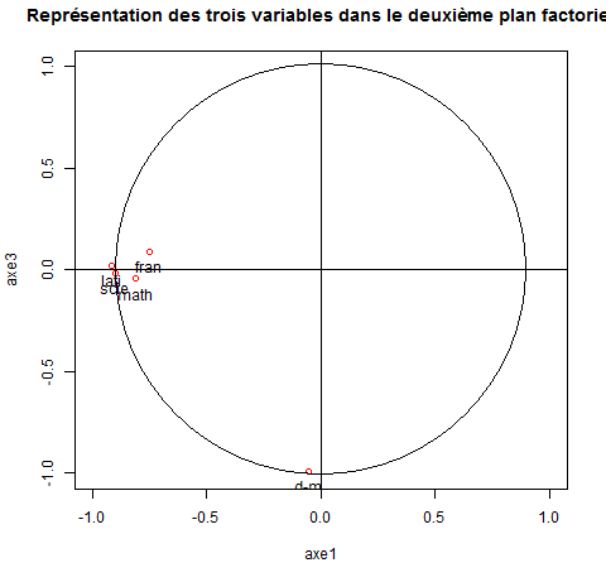
2.2 Utilisation des outils R

Q1) Effectuer l'ACP du jeu de données notes étudiées 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 :







Q2) 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°.

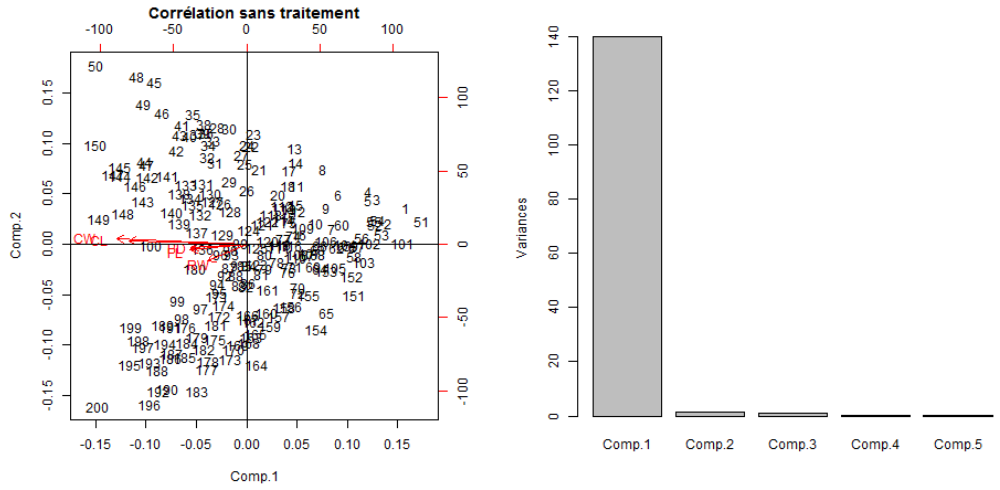
La fonction *biplot.princomp* donne accès à des options supplémentaires par rapport à la fonction *biplot*. En argument nous avons l'objet de la classe *princomp*, nous avons aussi la valeur *choices* pour définir la taille des vecteurs pour le plot. Nous avons aussi la valeur *scale* pour obtenir une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

Comme dans l'exercice 1, on s'intéressera aux données crabs, et plus particulièrement aux descripteurs quantitatifs. On commence donc par charger les données et sélectionner les variables quantitatives.

Q1) 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 confirmé dans les questions précédentes.

Q2) 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 séparé la variable *CL* et avons divisé le reste des varible par le vecteur ainsi obtenu. Ce qui nous permet de nous affranchir de l'influence de la taille général du crabe. Nous avons ensuite refait une ACP et nous avons représenté les données obtenues.

les variables ne sont ainsi plus corrélées comme précédemment ce qui nous donne les graphiques suivants :

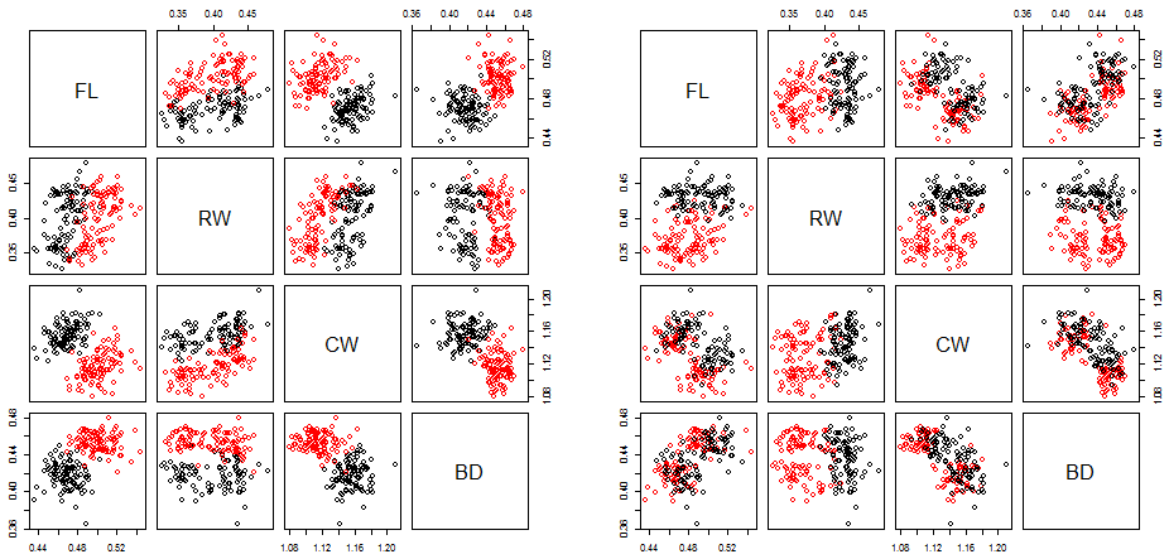


Figure 1: A gauche, graphique permettant de différencier les espèces. A droite, graphique permettant de différencier les sexes des crabes.

Conclusion :

Effectuer un traitement préalable des données crabs nous a permis de mettre en évidence les différences significatives existantes pour chacun des critères en fonction du sexe et de l'espèce de l'individu. Cela peut par exemple permettre de déterminer ces deux caractéristiques (sexe et espèce) en fonction de la mesure des 4 paramètres de taille retenus.
On comprend alors l'utilité d'un tel traitement et l'exploitation que l'on peut faire des données pour peu qu'on sache extraire l'information.
De manière générale, ce TP nous a permis d'exploiter des jeux de données et d'en extraire des informations grâce à différents outils. Il a mis en évidence la complexité d'interpréter un grand nombre de données et la nécessité de parvenir à une représentation concluante et explicite.