# **SY09 Printemps 2009**

# **Statistique descriptive, Analyse en composantes principales**

## Exercice I : Statistique descriptive

#### Exercice I.1 : Données babies

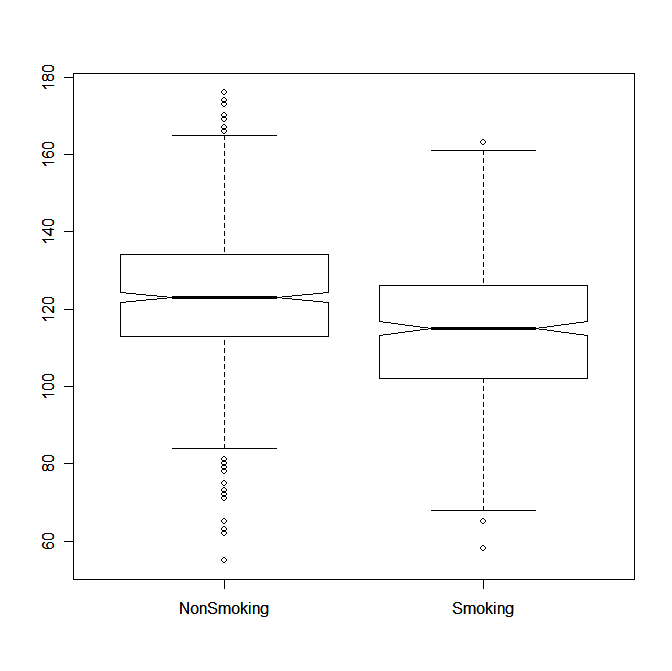
Cet exercice consiste à analyser un tableau de données constitué de 1236 individus (bébés) décrits au travers de 23 variables. L’analyse effectuée ne s’intéressera qu’à 8 de ces variables.

##### Questions

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

L’objectif ici est de mettre en relation des variables de différent types (qualitative et quantitative) et de chercher à mettre en lumière les liens entre ces variables.

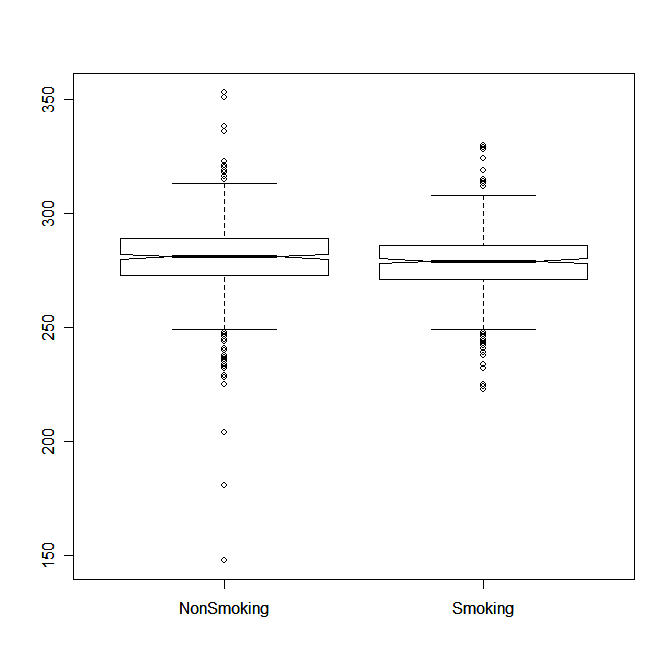
Nous avons opté pour une représentation de type “boite à moustache” affichant les intervalles de confiance pour chaque distribution (fonction boxplot, avec le paramètre notch à True ). En comparant les graphes relatifs aux mères fumeuses et non fumeuses, il est possible de déterminer, à l’aide des intervalles de confiance, l’existence ou non d’une relation entre le fait de fumer et le poids du bébé à la naissance.



On remarque en effet que les intervalles de confiance ne se chevauchent pas (intersection vide). On peut donc conclure que le fait que la mère soit fumeuse ou non a une influence significative sur le poids du bébé à sa naissance.

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

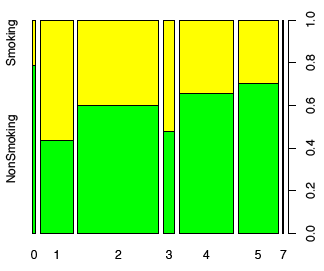
Cette question est similaire à la précédente (mise en relation de variables quantitatives et qualitatives), nous avons donc opté pour la même méthode (graphes en boite à moustache avec intervalle de confiance).



Contrairement au résultat de la question 1, les intervalles de confiance se chevauchent. Ces résultats ne nous permettent pas d’affirmer de manière significative que le temps de gestation est influencé par le fait de fumer ou non.

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

Comparant ici uniquement des variables qualitatives, nous avons opté pour un histogramme, ce dernier mettant en valeur le nombre d’échantillons pour chaque valeur possible des variables.



Cet histogramme représente le taux de mères fumeuses ou non, en fonction de leur niveau d’études. Chaque barre verticale représente un niveau d’étude donné, la largeur des barres indique le nombre d’échantillons disponibles pour chaque niveau d’étude. Enfin, les zones colorées représentent les proportions de fumeuses (jaune) et non fumeuses (vert).

Au vu des quantités d’échantillons trop faibles (relativement aux autres), il nous semble justifié de ne pas tenir compte des résultats pour les niveaux 0 (less than 8th grade) et 7 (Trade school HS unclear).

On remarque que le nombre de mères non fumeuses est proportionnel au niveau d’étude de ces dernières. Ainsi, il existe bel et bien une relation entre le niveau d’étude et le fait du fumer ou non.

### Conclusion : Rapprochement avec l’article extrait de l’édition du New York Times du 1er mars 1995

L’article du NY Times rapporte une analyse sur la mortalité à la naissance. Il explique les morts sont principalement dues à un temps de gestation trop court, et que le poids à la naissance est, au contraire, un élément non déterminant.

Notre analyse ayant montré que la consommation de cigarettes par la mère durant la grossesse influe seulement sur le poids du bébé à la naissance, et pas sur le temps de gestation. Nos conclusions sont donc similaires à celle de l’article, c’est-à-dire qu’une mère fumeuse ne met pas son futur enfant plus en danger qu’une non fumeuse.

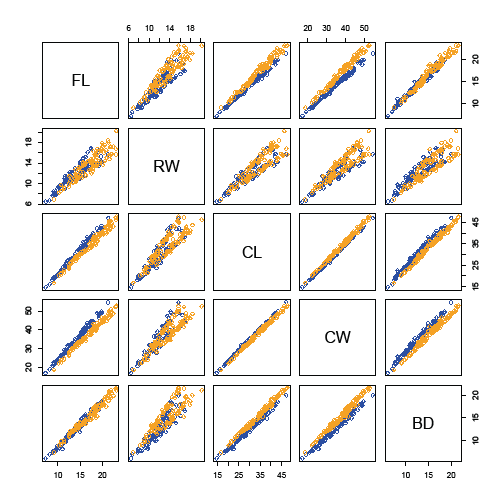
#### Exercice I.2 : Données crabs

On considère un jeu de données constitué de 200 individus (crabs) décrits par 8 variables. On ne considérera que les variables quantitatives (au nombre de 5).

##### Questions

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 de ces caractéristiques ?*





Ces graphiques realisés grace à la commande pairs() de R repéresentent les différentes variables quantitatives les unes en fonction des autres. Le premier graphique met en evidence la variables qualitatitives sexe (rouge et bleu) , le deuxième l’ espece (jaune et bleu).

Malgré de très fortes similitudes, nous pouvons tout de même remarquer des tendances différentes en ce qui concerne le sexe. Il semble donc possible d’identifier le sexe d’un crabe suivant une serie de mesures données. Cepdendant ces resultats ne peuvent être considérés comme significatifs. Les resultats obtenus en ce qui concerne l’espece ne semblent clairement pas montrer de differences significatives dans les mesures.

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 ?*

D’après les graphiques obtenus précédemment, on remarque que toutes les données sont fonctions approximativement lineaires les unes des autres. Ceci dégage une rélle corrélation entre les différents jeux de données. Nous avons, pour confirmer cette interdependence entre les variables quantitatives, effectuer le calcul du coefficient de correlation entre les différentes variables :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | BD | CW | CL | RW |
| FL | 98.7% | 96.5% | 97.8% | 90.7% |
| RW | 89% | 90% | 89.3% |  |
| CL | 98.3% | 99.5% |  |  |
| CW | 96.8% |  |  |  |

Ce tableau représente les coefficients de corrélation entre les différentes variables étudiées. Tous ces coefficients sont très élevés, ce qui confirme la très forte interdépendance de toutes les variables quantitatives. Afin de palier à ce problème, il est nécéssaire de réduire le problème aux variables les plus significatives, et donc de procéder à une analyse par composantes principales.

## Exercice II : Analyse en composantes principales

### **Exercice II.1 : Exercice théorique**

La matrice utilisée pour cet ACP théorique contient 4 individus et 3 variables (n = 4 et p = 3).

Matrice A :

**1)** **Afin de réaliser l’ACP des données, nous avons :**

* Centréla matrice de départ en colonne : (Ac <- A – matrix(1,4,1) %\*% apply(A,2,mean) ).

Matrice A centrée :

* Calculé la matrice de variance S : ( S <-1/4 \* t(Ac) %\*% Ac ).

Matrice de variance S :

* Réalisé l’analyse spectrale de S (valeurs/vecteurs propres grâce à la fonction eigen(S) de R)

Matrice de variance S :

On obtient ainsi la matrice U des vecteurs propres ordonnés selon leurs valeurs propres décroissantes (sous R) : associé aux valeurs propres 2, 1 et 0.5.

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

axe 1 : x 100 = 57.14 %

axe 2 : x 100 = 28.57 %

axe 3 : x 100 = 14.29 %

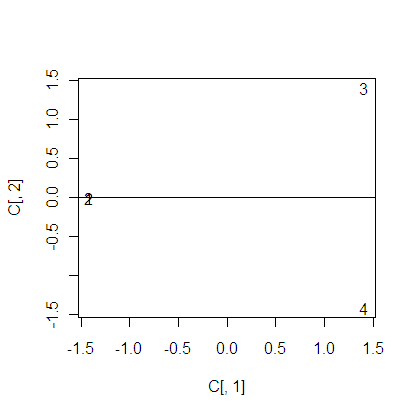
C’est l’axe 3 qui possède le pourcentage d’inertie le plus faible, nous nous intéresserons donc moins à cet axe.

**2) Calcul des composantes principales :**

Nous calculons les composantes principales de A en projetant cette dernière sur la base des vecteurs propres (matrice U).

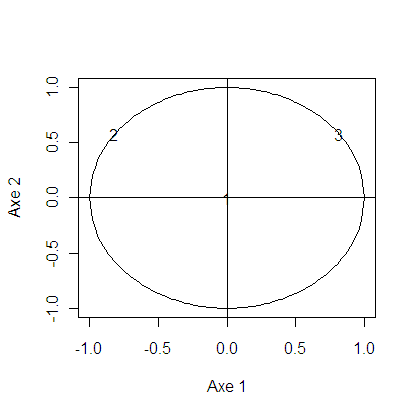
C = A %\*% U =

Représentation des quatre individus dans le premier plan factoriel : c'est-à-dire avec l’axe1 et l’axe 2 (les deux premières colonnes de C).



Représentation des 4 individus dans le premier plan factoriel. (Les individus 1 et 2 sont confondus et les individus 3 et 4 sont symétriques par rapport au deuxième axe factoriel).

**3) Représentation des trois variables dans le premier plan factoriel :** on utilise la formule D = D1/σUL0.5



Représentation des 3 variables dans le premier plan factoriel. (La variable 1 est au centre, donc représenté par aucun des axes, les variables 2 et 3 sont symétriques par rapport au premier axe factoriel et situé sur le cercle).

**4) Calcul de l’expression  σ u’σ**

cσ u’σ :  🡪  σ u’σ**:**  🡪  σ u’σ**:** = A centrée

On observe une reconstruction progressive, colonne par colonne, de la matrice de départ A lorsque k augmente de 1 vers 3 jusqu'à retrouver la matrice A.

### Exercice II.2 : Traitements des données Crabs

Le but de l’exercice est d’utiliser l’ACP pour trouver une représentation des crabes qui permettent de distinguer visuellement différents groupes, liés à l’espèce et au sexe.

#### ACP sans traitement préalable

Après application de l’ACP et la représentation des données, on se retrouve confontré au problème

L’application d’une ACP directement sur le jeu de données brut fait apparaître un problème courant en analyse de données : **l’effet de taille**.

En effet, on remarque que le résultat de l’ACP nous donne un premier axe factoriel expliquant 98% du jeu de données. En représentant graphiquement ces données, on s’aperçoit qu’il y a clairement proportionnalité entre les caractéristiques quantitatives des différents individus. Un crabe ayant un lobe frontal grand aura en toute logique un corps proportionné en conséquence. Il existe donc une certaine proportionnalité entre l’attribut A d’un petit crabe et l’attribut A d’un gros crabe, de même pour les autres attributs. On le voit sur le graphique suivant obtenu grâce à la fonction « pairs » :

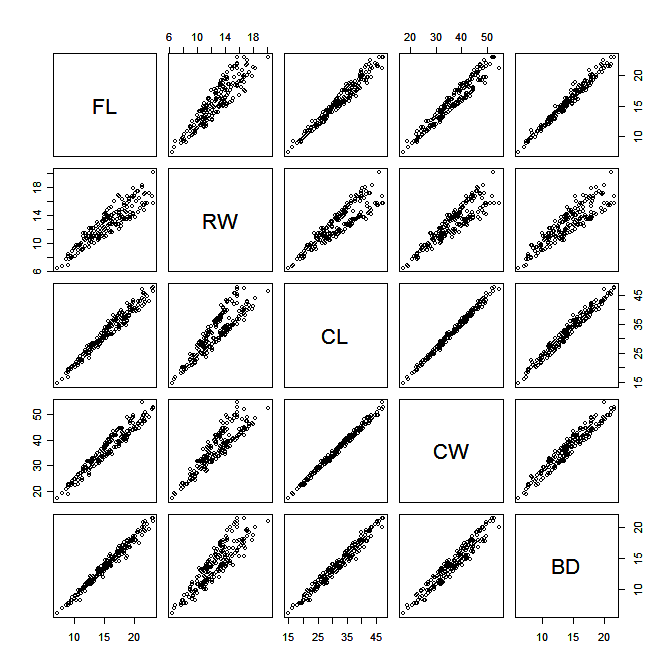


Figure : Représentation des variables quantitatives

On remarque des lignes de tendance très fortes dans chaque cas.

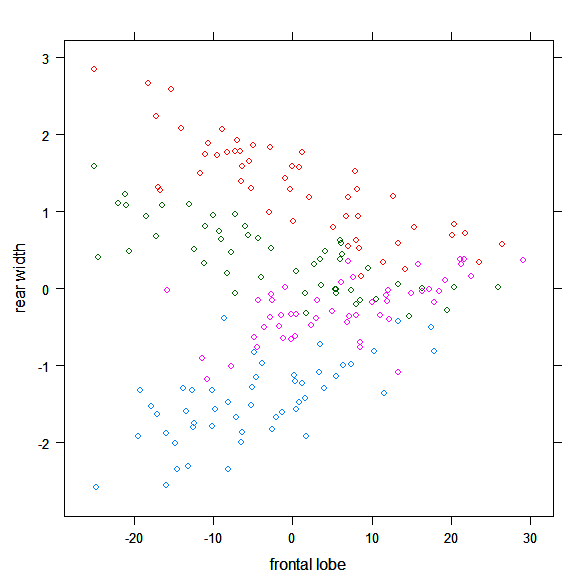


Figure : Représentation graphique dans le premier plan factoriel

Il faut donc trouver une solution pour améliorer la représentation graphique.

#### ACP avec traitement préalable

La solution pour améliorer la qualité de notre représentation en termes de visualisation des différents groupes est d’homogénéiser les données. Pour cela, il faut prendre une des variables puis diviser les autres variables par cette variable. Le facteur de proportionnalité qui existe entre les 5 valeurs d’un même crabe disparait.

Dans notre cas, nous avons divisé les autres variables avec les valeurs de la variable « longueur de la carapace ». On remarque tout de suite en calculant les inerties expliquées cumulés que la représentation est meilleure. Les deux premiers axes factoriels décrivent maintenant 88%, ce qui est très satisfaisant.

La différence se voit aussi sur la représentation graphique de toutes les variables obtenues avec la fonction « pairs » :

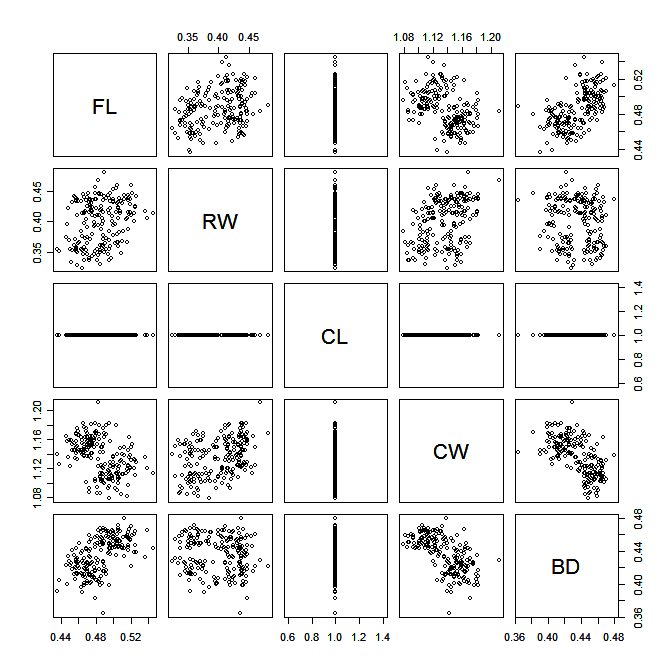
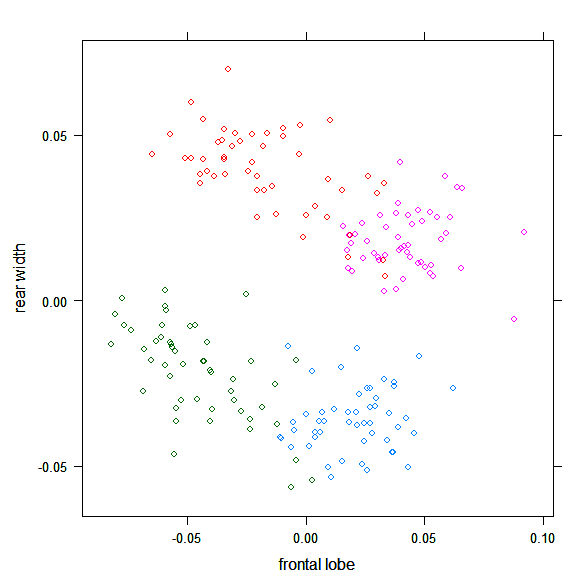


Figure : Représentation graphique des données après traitement

On remarque que les lignes de tendance ont disparu et des nuages de point ont apparu.



Grâce à ces modifications, la représentation dans le premier plan factoriel permet de clairement distinguer et l’espèce et le sexe d’un crabe donné.