**SY09 Printemps 2009**

**Statistique descriptive, Analyse en composantes principales**

Exercice I : Statistique descriptive

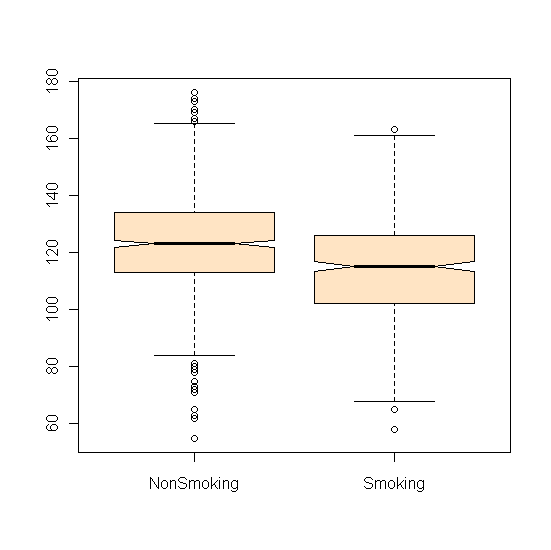
On part d’un jeu de donnée (de type « tableau individu/variable ») constitué de 1236 bébés décrits par 23 variables. Pour la réalisation de cet exercice, nous travaillerons sur 8 des 23 variables.

***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 ?*

On cherche à mettre en évidence l’effet et/ou les liens d’une variable qualitative (mère fumeuse ou non pendant la grossesse) sur une variable quantitative (le poids de naissance).

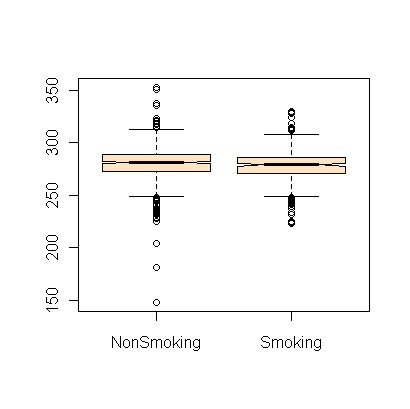
Je me suis tout d’abord penché sur le calcul des moyennes des poids de naissance pour les deux types (fumeurs et non fumeurs). La représentation sous forme d’histogramme n’apportant rien, j’ai réalisé une boîte à moustache, en y faisant figurer les intervalles de confiance.



*Boîte à moustaches* *Tableau récapitulatif*

Les intervalles de confiance étant disjoints, il apparaît ainsi 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, comme le laisse penser la différence d’environ 8 oz entre les moyennes.

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 ?*

Il s’agit également de comparer l’effet d’une variable qualitative sur une variable quantitative. J’utilise donc encore une boîte à moustaches, car cette méthode est plus comparative, après calcul des moyennes.

Je n’observe pas de différence réellement significative entre la durée de gestation des mères fumeuses et celle des mères non-fumeuses. On ne peut donc affirmer qu’une mère fumant pendant sa grossesse aura un temps de gestation inférieur à une mère qui ne fume pas.

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

Il s’agit maintenant de comparer l’effet d’une variable qualitative (le niveau d’étude) sur une autre variable qualitative (mère fumeuse ou non pendant la grossesse).

Dans un premier temps, j’ai réalisé la moyenne du niveau d’études des mères fumeuses puis celle des non-fumeuses (omission de la ligne « babies$education<-factor(babies$education,ordered=T » de l’énoncé).

Moyenne niveau d’éducation des mères fumeuses = 2.69

Moyenne niveau d’éducation des mères fumeuses = 3.08

Cependant, le niveau d’éducation étant une variable qualitative, ces données sont inexploitables.

J’ai donc réalisé un tableau de contingence afin de croiser les modalités d’une variable avec celles de l’autre. Je divise ensuite chaque cellule par la somme de chaque ligne correspondante. Le tableau va ainsi indiquer l’influence du niveau d’étude sur le fait que la mère soit fumeuse.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Non-fumeurs | Fumeurs | Somme des lignes |
| 0 | 15 | 4 | 19 |
| 1 | 79 | 102 | 181 |
| 2 | 264 | 176 | 440 |
| 3 | 30 | 33 | 63 |
| 4 | 194 | 102 | 296 |
| 5 | 154 | 65 | 219 |
| 7 | 6 | 1 | 7 |
| 9 | 0 | 1 | 1 |

|  |  |  |
| --- | --- | --- |
|  | % de Non-fumeurs par niveau d’étude | % de fumeurs par niveau d’étude |
| 0 | 78% | 22% |
| 1 | 43% | 57% |
| 2 | 60% | 40% |
| 3 | 48% | 52% |
| 4 | 66% | 34% |
| 5 | 71% | 29% |
| 7 | 86% | 14% |
| 9 | 0% | 100% |

Une différence significative apparaît donc après 4 ans d’étude : Plus le niveau d’étude est haut, plus le pourcentage des femmes non-fumeurs est important. On suppose donc que le niveau d’étude influe sur le fait que la mère soit fumeuse. Cependant, la taille des échantillons aux niveaux 7 et 9 étant trop faible, on ne peut appuyer notre analyse sur ces pourcentages. De même, pour les niveaux 1 et 3, la différence n’est pas suffisamment important pour affirmer qu’à toux les niveaux d’études, cette variable réduit les chances qu’une mère soit fumeuse.

Voir pour histogramme et si nécessaire lien avec article !

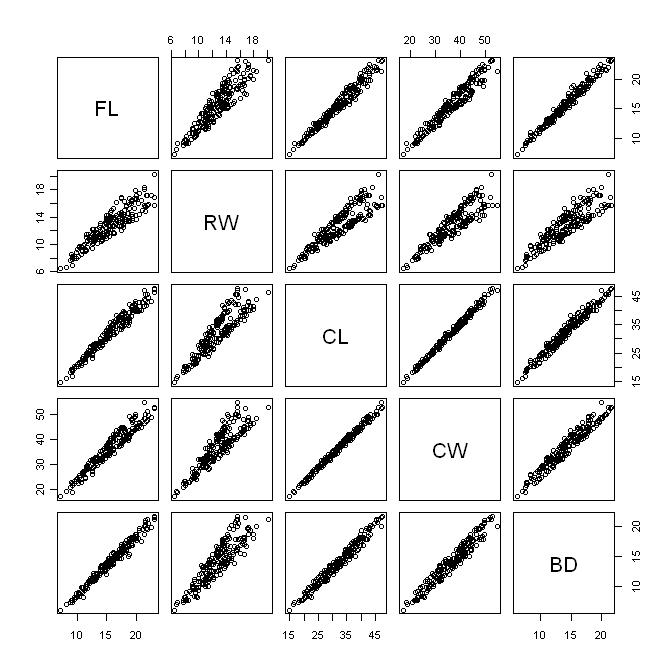
Exercice II : Analyse en composantes principales

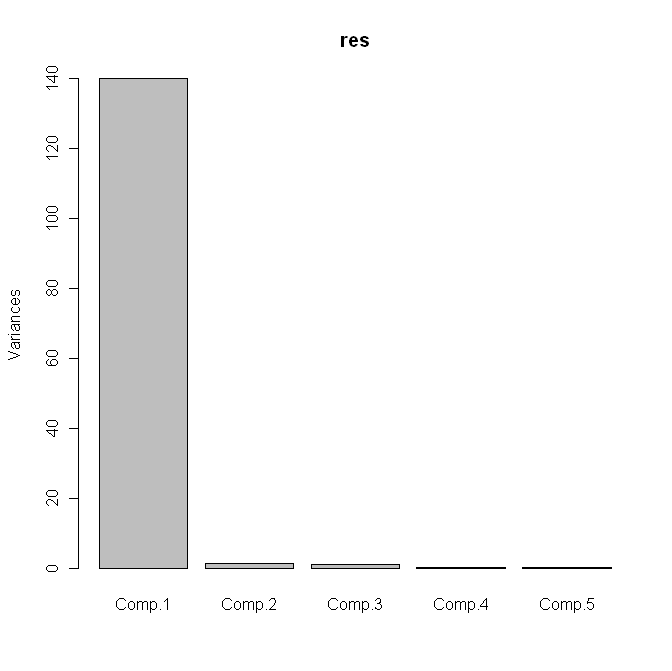
**Exercice II.1 : Exercice théorique**

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

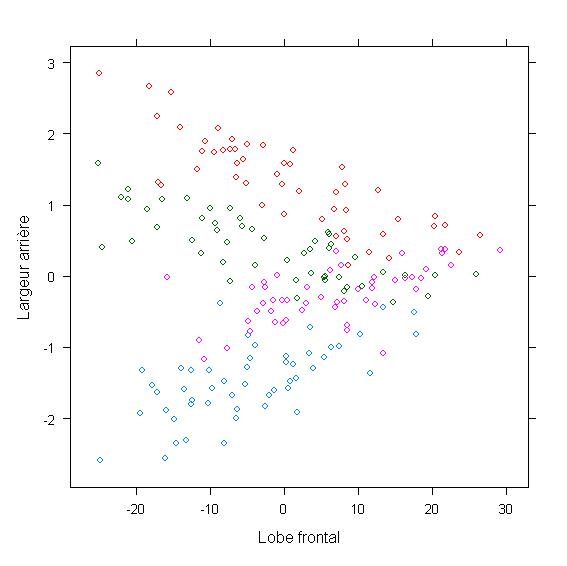
On cherche dans cette dernière partie à trouver une représentation des données permettant de distinguer les crabes selon leurs espèces et leurs sexes. Ce traitement va se baser sur une utilisation de la méthode ACP.

1. **ACP sans traitement préalable**

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**.



Le premier axe factoriel explique 98% du jeu de données. La représentation graphique des données montre qu’il y a une proportionnalité entre les caractéristiques quantitatives des différents individus. Il y a proportionnalité entre la composante A d’un petit crabe et la composant A d’un gros crabe.



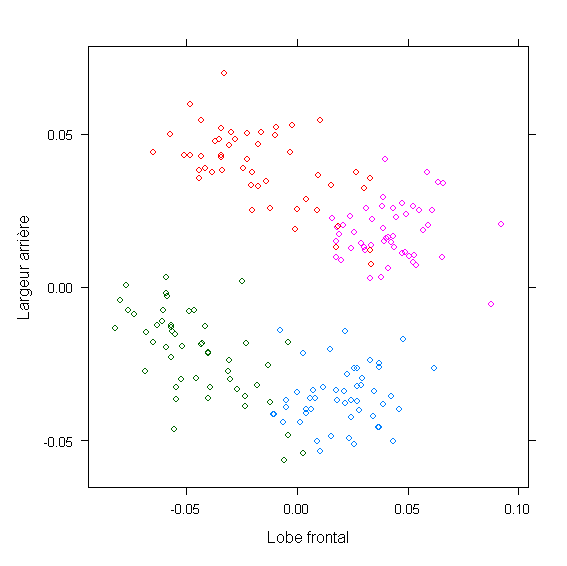
Cependant, la représentation graphique dans le premier plan factoriel ne nous permet pas de grouper graphiquement les différentes populations.

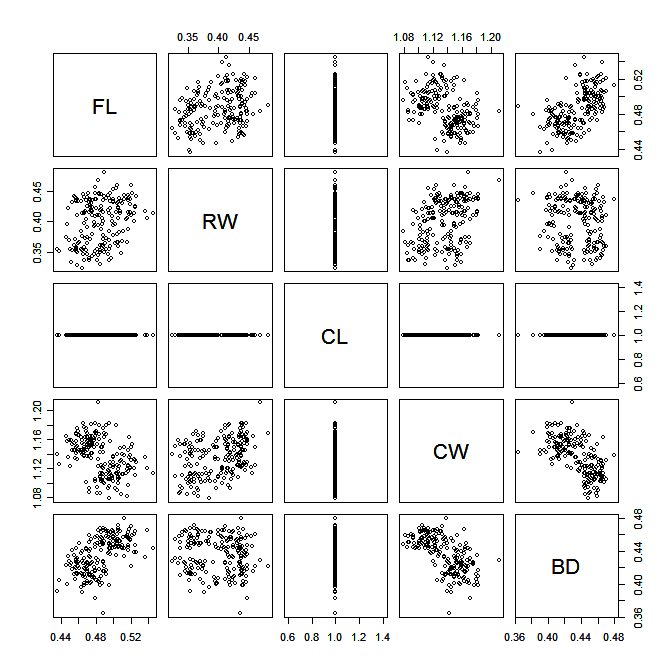
Il faut donc réaliser un pré-traitement pour améliorer les résultats de l’ACP.

1. **ACP avec traitement préalable**

Pour réduire l’effet taille, il apparait judicieux d’homogénéiser les données en divisant les cinq variables des individus par une des variables de cet individu. Le facteur de proportionnalité qui existe entre les 5 valeurs d’un même crabe disparait.

J’ai donc divisé les autres variables avec les valeurs de la variable « longueur de la carapace ». Grâce aux inerties expliqués cumulés, il apparaît que la représentation est meilleure (les deux premiers axes factoriels décrivent maintenant 88%).

On obtient ainsi les représentations graphiques suivantes (fonction  « pairs » et représentation dans le premier plan factoriel.



Les lignes de tendance ont disparu et des nuages de points sont apparus.

La représentation dans le premier plan factoriel permet de clairement distinguer et l’espèce et le sexe d’un crabe donné.