**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 ?*

Nous voulons comparer **une variable quantitative** (poids de naissance) et **une variable qualitative** (mère fumeuse durant la grossesse) ; nous cherchons alors à décrire les liens entre les deux couples de variables.

Deux modes de représentation sont alors possibles :

* Une représentation sous forme d’histogramme : cependant, cela ne nous permet de conclure quant à une différence significative entre les deux couples de variables ; puisque les zones où évoluent nos histogramme restent assez proches.
* Une autre représentation possible est celle sous forme de « boite à moustaches » : grâce à la fonction « notch=T », nous pouvons faire figurer, sur les boites, les intervalles de confiances. Lorsque les 2 intervalles de confiance ne se chevauchent pas, alors la différence entre les 2 couples de variables est significative. Ce qui est le cas ici. Cette seconde représentation nous permet donc 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.

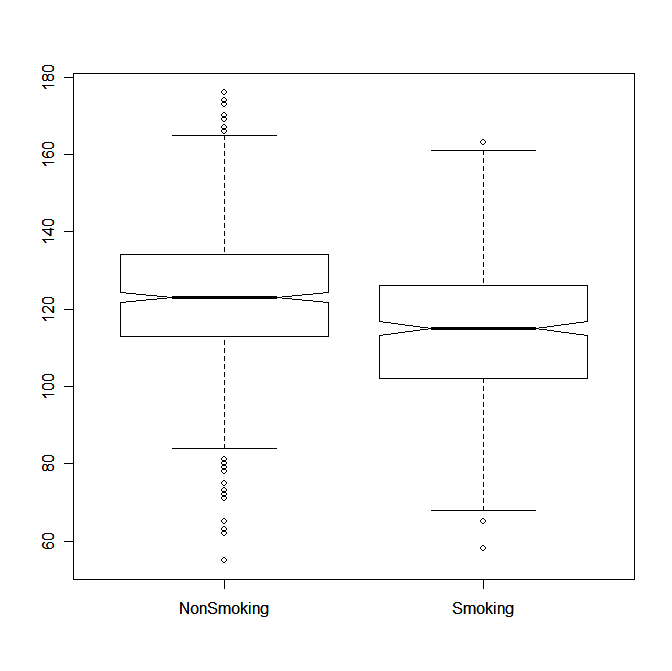


Figure : boite à moustaches

Ainsi, lorsqu’une mère fume durant la grossesse, le poids du bébé sera inférieur à celui d’un bébé naissant d’une mère non-fumeuse.

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 **une variable quantitative** (temps de gestation) et **une variable qualitative** (mère fumeuse durant la grossesse) ; nous cherchons à décrire les liens entre les deux couples de variables.

Nous optons donc pour le mode de représentation en boite à moustaches, plus comparatif, comme vu précédemment.

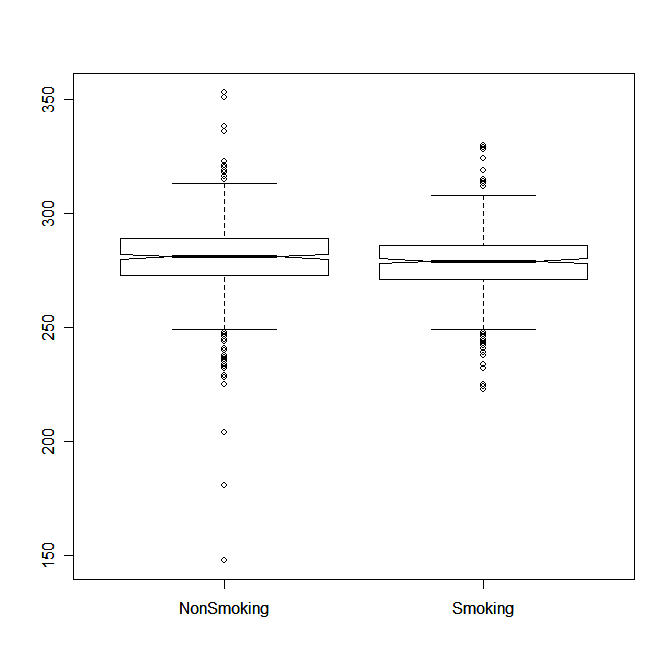


Figure : boite à moustaches

Ces résultats ne nous permettent pas d’observer une différence réellement significative cette fois, entre le temps de gestation d’une mère fumeuse durant la grossesse et d’une mère non fumeuse.

Ainsi, nous ne pouvons pas affirmer que lorsqu’une mère fume durant la grossesse, le temps de gestation sera inférieur à celui d’une mère non-fumeuse.

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

Il s’agit cette dois-ce de comparer **deux variables qualitatives** : si la mère est fumeuse, et le niveau d’étude de la mère.

Pour cela, nous commençons par réaliser un tableau de contingence, croisant ainsi les modalités d’une variable avec celles de l’autre (en colonne, si la mère est fumeuse ; en ligne, le niveau d’étude). Puis, nous faisons la somme de chaque ligne, pour ensuite diviser chaque cellule par le somme de la ligne correspondante. Ces calculs nous permettent alors d’observer l’**influence** qu’a le niveau d’étude sur le fait que la mère soit fumeuse.



Figure : Transformation du tableau de contingence

Ces résultats nous permettent d’observer une différence significative après 4 ans d’études. Les femmes fument moins quand elles ont un certain haut niveau d’étude alors qu’en dessous de 3 ans d’étude, il y a plus de femmes qui fument. La première et la dernière ligne du tableau ne sont pas à prendre en compte à cause du faible nombre de personnes étudiées.

Ainsi, le niveau d’étude a une influence sur le fait que la mère soit fumeuse ou non.

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

* « low birth weights per se are the cause of the high infant mortality rate in the United Sates »
* « being born too soon […] is the main underlying cause of stillbirth and infant deaths within four weeks of birth »

Or, notre première étude comparative a montré que si la mère fumait durant la grossesse, le poids du bébé serait inférieur à celui obtenus pour une mère non fumeuse. Fumer durant la grossesse nuirait donc à la survie de l’enfant car un faible poids de naissance serait la cause d’un taux élevé de mortalité infantile aux Etats-Unis.

Cependant, le texte souligne le paradoxe suivant : « the babies of smoking mother had a higher survival rate […], smoking interferes with weight gain, it does not shorten pregnancy ». Une mère fumeuse donnerait alors naissance à un bébé au poids plus faible, mais n’aurait pas une grossesse plus courte. Ce qui se rapproche des résultats obtenus dans la question 2, puisque nous n’avons pas réussi à observer que le temps de gestation serait plus court pour une mère fumeuse.

En conclusion, nos résultats confirmeraient les données de l’article.

Exercice II : Analyse en composantes principales

**Exercice II.1 : Exercice théorique**

1. Nous partons d’une matrice A, à 4 individus et 3 variables, et nous souhaitons lui appliqué la méthode d’analyse en composantes principales. Pour l’application de cette méthode, la matrice A doit être **centrée en colonne**. Nous centrons donc la matrice, puis nous calculons ensuite la **matrice de variance S**. La fonction « eigen(S) » nous permet alors de récupérer les matrices des **vecteurs et des valeurs propres**. On a donc :

Matrice A : Matrice A centrée : Matrice de variance S :

Matrice des vecteurs propres U : associé aux valeurs propres 2, 1 et 0.5

**Calcul des pourcentages d’inertie expliq**uée par chacun de ces axes :

axe 1 : x 100 = 57.14 %

axe 2 : x 100 = 28.57 %

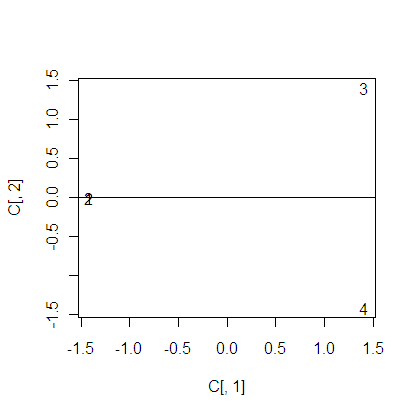
axe 3 : x 100 = 14.29 %

(L’axe 3 est ici moins intéressant que les deux premiers, puisqu’il explique seulement 14% de l’inertie générale)

1. **Calcul des composantes principales :** celles-ci s’obtiennent aisément grâce à la formule : C= AU, où A est la matrice centrée en colonnes, et U la matrice des vecteurs propres. Ainsi :

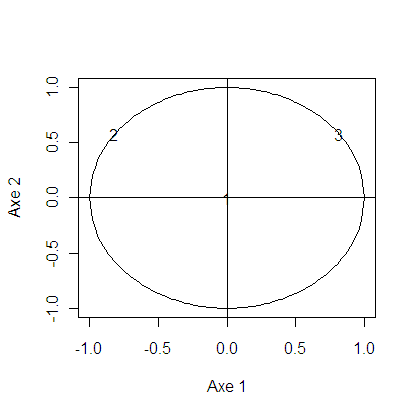
Matrice des composantes principales C :

**Représentation des quatre individus dans le premier plan factoriel :** Le premier plan factoriel est ∆u1, composé des u1 et u2 de la matrice C.



On observe que les deux premiers individus sont confondus sur les 2 axes, et que les deux derniers individus sont similaires sur l’axe 1 mais éloignés sur l’axe 2. Le premier axe permet donc d’isoler les 2 premiers individus aux 2 derniers. Sur le second axe permet quant à lui de différencier les 2 derniers individus (les 2 premiers individus restant similaires).

1. **Représentation des trois variables dans le premier plan factoriel :** Nous obtenons la représentation grâce à la formule : D = D1/σUL0.5



On observe que les variables 2 et 3 sont similaires sur l’axe 2 et qu’elles sont bien représentées (proche du cercle). Quant à la 1ère variable, elle est mal représentée et est moins similaires aves les 2 dernières variables, 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.

1. **Calcul de l’expression  σ u’σ**

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

**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.

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**.

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 » :

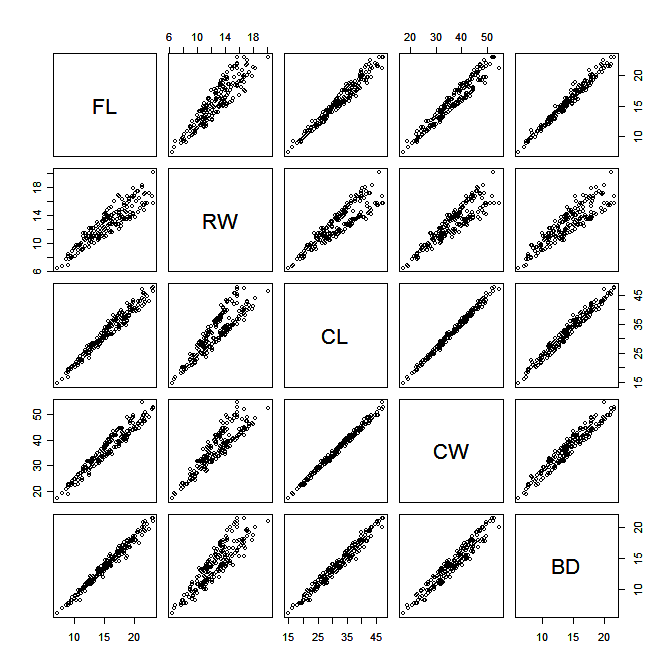


Figure : Représentation des variables quantitatives

On remarque des lignes de tendance très fortes dans tous les cas.

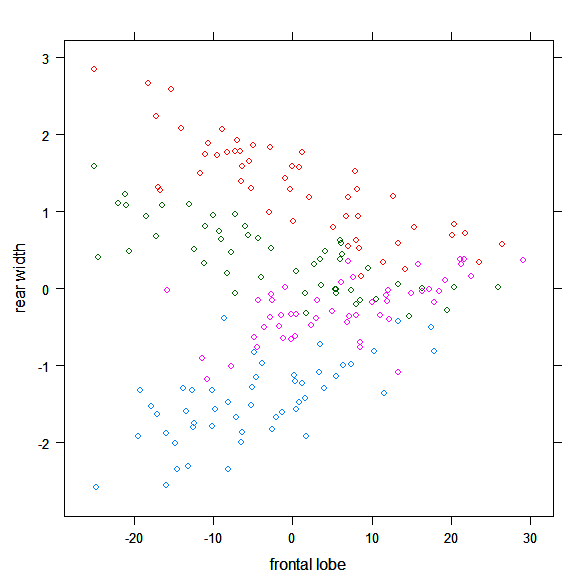


Figure : Représentation graphique dans le premier plan factoriel

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

1. **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 » :

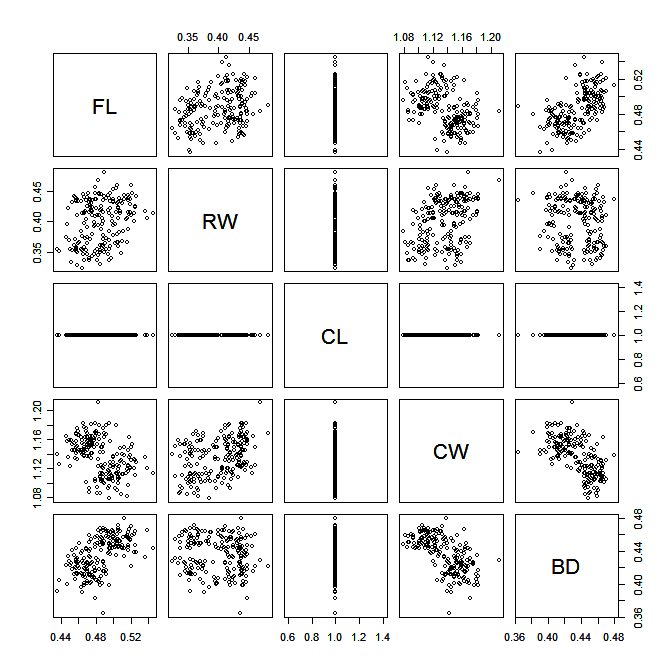
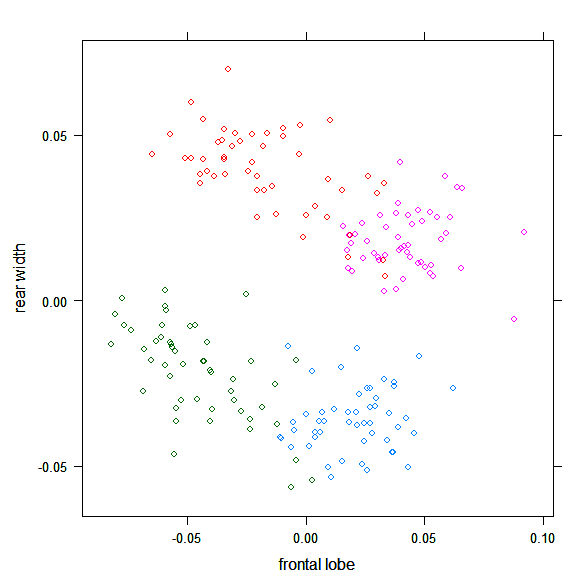


Figure : Représentation graphique des données après traitement

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



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