# SY09 - TD01

Pierre-Alexandre Fonta, Perrine Letellier

3 avril 2013

L'objectif de cet exercice est de traiter une grande quantité de données par les outils de statistiques descriptives élémentaires afin de mener une analyse.

## 1 Statistique descriptive

### 1.1 Données babies

Soit un jeu de données sous forme d'un tableau individus-variables ayant en entrées une population de 1236 bébés décrits par 13 variables. Seules 8 variables sont conservées pour l'analyse voulue : 5 quantitatives (poids à la naissance, durée de gestation, nombre de grossesses précédentes, taille de la mère, poids de la mère) et 3 qualitatives (âge de la mère, mère fumeuse ou non, 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 ?**
Étant données deux populations d'enfants (nés d'une mère fumeuse ou non) à comparer selon leur poids à la naissance (la variable "bwt"), nous commençons tout d'abord par étudier un résumé numérique de chacune des population :

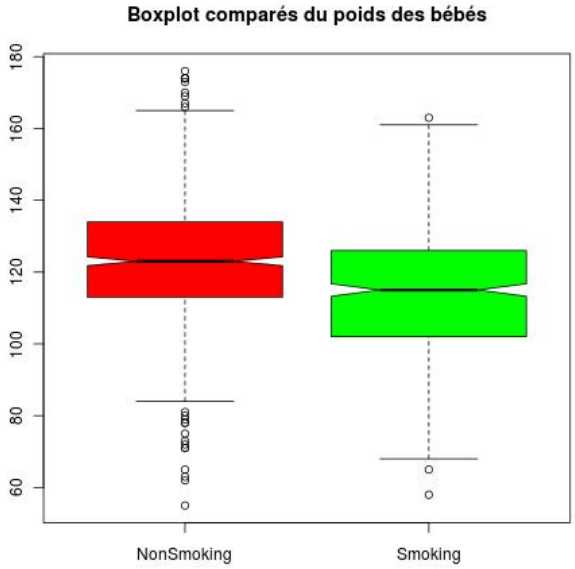
| Min. | 1st. Qu. | Med. | Mean | 3rd. Qu. | Max |
|---|---|---|---|---|---|
| 58.0 | 102.0 | 115.0 | 114.1 | 126.0 | 163.0 |

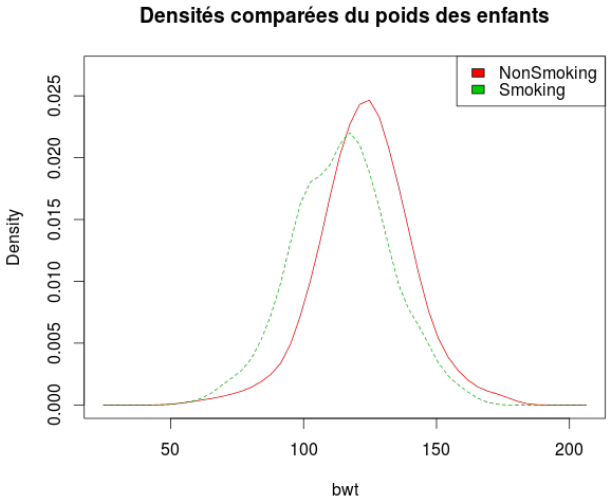
TABLE 1 – Résumé numérique pour le poids des enfants nés de mères fumeuses

| Min. | 1st. Qu. | Med. | Mean | 3rd. Qu. | Max |
|---|---|---|---|---|---|
| 55.0 | 113.0 | 123.0 | 123.0 | 134.0 | 176.0 |

TABLE 2 – Résumé numérique pour le poids des enfants nés de mères non-fumeuses

Cet apercu des données, et notamment si l'on s'attache aux statistiques de base que sont la moyenne et la médiane, nous permet de dégager une différence entre les deux populations. Les enfants de mères fumeuses semblent avoir tendance à naître plus maigres que ceux de mères non fumeuses. Avec un écart de 8 points sur la médiane et de 9 sur la moyenne. On peut vérifier cette tendance par une visualisation sous forme de box-plot, ainsi que l'observation de la densité des deux populations :




Les densités montrent ce que l'on observait numériquement : elles ne sont pas centrées. L'approche box plot confirme également la tendance : pour chaque population nous avons identifié la médiane. Les intervalles de confiance sur chaque médiane ne se chevauchent pas, la différence de poids est donc statistiquement significative dans 95% des cas.

On peut enfin vérifier la non égalité des moyennes en appliquant un test bilatéral de Student sur les deux populations : nous considérons ces populations indépendantes et normalement distribuées. La p-value du test de Student est de $2{,}2.10^{-16} \leqslant \alpha = 0{,}05$ , nous rejetons

définitivement l'hypothèse d'égalité de poids. **Un bébé de mère fumeuse est donc statistiquement plus maigre qu'un bébé de mère non fumeuse.**

**Est-ce qu'une mère qui fume pendant sa grossesse est encline à avoir un temps de gestation plus court qu'une mère qui ne fume pas ?**
De même que précédemment, nous commençons par un résumé numérique de chaque population selon la variable "gestation" :

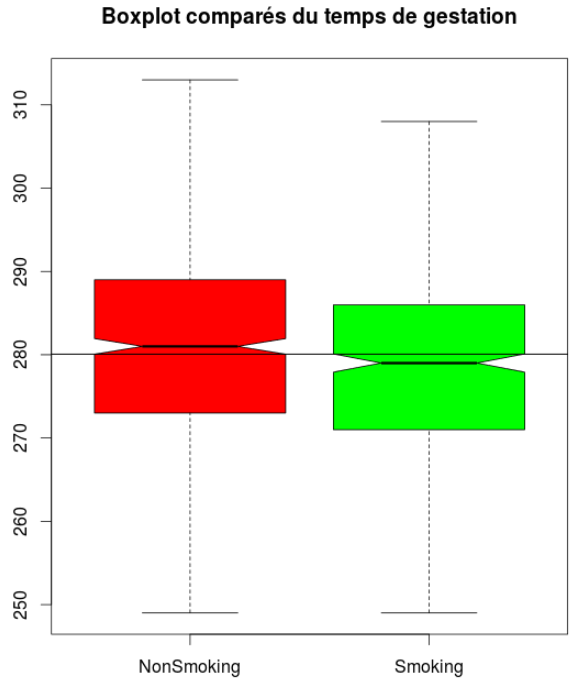
| Min. | 1st. Qu. | Med. | Mean | 3rd. Qu. | Max |
|---|---|---|---|---|---|
| 223.0 | 271.0 | 279.0 | 278.0 | 286.0 | 330.0 |

TABLE 3 – Résumé numérique pour le temps de gestation des mères fumeuses

| Min. | 1st. Qu. | Med. | Mean | 3rd. Qu. | Max |
|---|---|---|---|---|---|
| 148.0 | 273.0 | 281.0 | 280.2 | 289.0 | 353.0 |

TABLE 4 – Résumé numérique pour le temps de gestation des mères non-fumeuses

A priori, les médianes sont légèrement différentes et on peut penser que le fait de fumer influe sur la gestation moyenne (2 points d'écart sur la médiane et la moyenne). Il est cependant impossible de conclure à ce niveau : la différence n'est pas signifiante. Nous observons donc les populations sous forme de box-plot. Les valeurs atypiques sont beaucoup plus nombreuses chez les femmes ne fumant pas, mais afin de voir plus particulièrement les intervalles de confiances sur les médianes, nous ne les affichons pas :




Les intervalles de confiance des deux médianes se chevauchent, nous ne pouvons donc pas affirmer que les mères fumeuses ont un temps de gestation plus court que les mères non fumeuses.

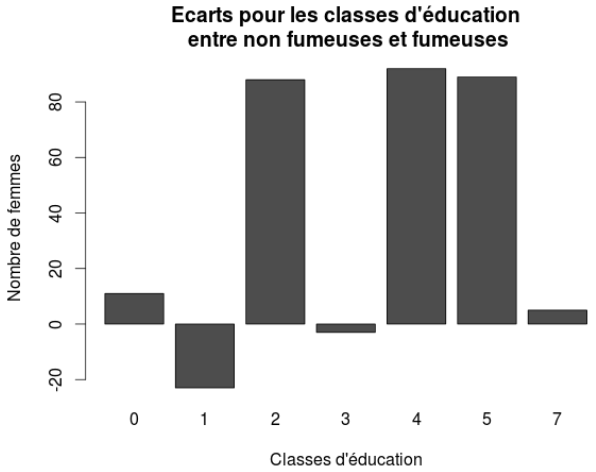
Si nous appliquons un test de Student aux deux populations, avec pour hypothèse de départ l'égalité entre le temps moyen de gestation pour les fumeuses et les non fumeuses, la statistique de retour t = -2.394 appartient à l'intervalle de confiance (à 95%) [-4.017; -0.398]. L'hypothèse ne peut donc pas être rejetée. L'écart temporel entre les deux durées de gestation n'est pas suffisamment grand pour être signifiant. **Statistiquement, nous ne pouvons pas affirmer qu'une mère qui fume a un temps de gestation plus court qu'une qui ne fume pas.**

**Le niveau d'études a-t-il une influence sur le fait que la mère soit fumeuse ?**
Il s'agit ici de comparer l'influence d'une variable qualitative "niveau d'étude" sur une autre "smoke", ce qui rend impossible la comparaison numérique telle qu'elle à été mise en oeuvre précédemment. Nous allons tester l'indépendance des ces deux variables. Auparavant, nous pouvons observer une tableau de contingence des populations par niveau d'études :

| | 0 | 1 | 2 | 3 | 4 | 5 | 7 |
|---|---|---|---|---|---|---|---|
| Smoking | 4 | 102 | 176 | 33 | 102 | 65 | 1 |
| NonSmoking | 15 | 79 | 264 | 30 | 194 | 154 | 6 |

TABLE 5 – Tableau de contingence

Grâce à ce tableau, on observe de grandes différences pour chaque classe d'étude entre le nombre de fumeuses et de non fumeuses (outre le fait qu'aucune mère des deux populations n'est représentée dans la classe 6). Pour les classes 2, 4 et 5, les différences sont les plus grandes : il y a jusqu'à 50% de non fumeuses en plus.



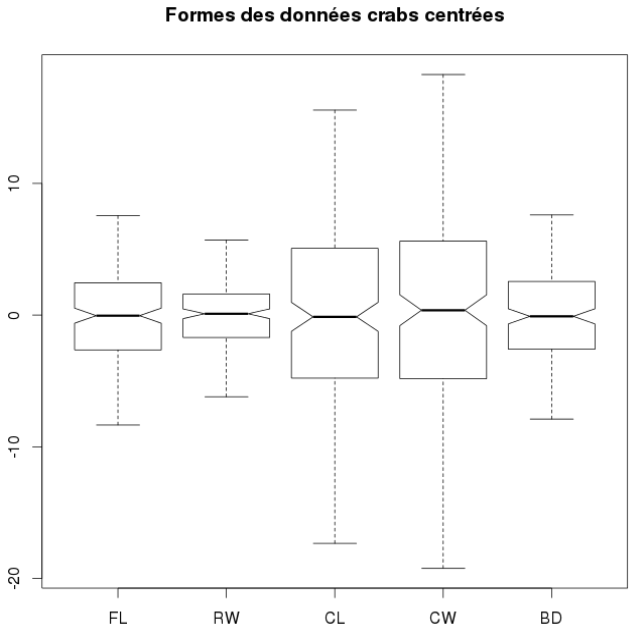

Cependant, selon les répartitions par classes, nous ne pouvons soutenir une hypothèse de distribution uniforme ni d'indépendance des deux variables. Nous effectuons donc un test du $\chi^2$ pour soutenir ou nous les affirmations précédentes. Selon l'hypothèse d'indépendance des variables "smoke" et "niveau d'étude", la p-value est de $1{,}459.10^7 \leqslant \alpha = 0{,}05$. Le seuil de basculement de $H_1$ vers $H_0$ est faible, Nous pouvons donc rejeter l'hypothèse d'indépendance, ce qui confirme nos observations. **Statistiquement, le niveau d'étude et le fait que la mère soit fumeuse sont corrélés.**

Nous avons tout d'abord affirmé la différence de poids entre un bébé de mère fumeuse et un autre de mère non fumeuse. Il a ensuite été impossible de déterminuer statistiquement l'influence du tabagisme sur le temps de gestation des mères. L'article de référence (extrait du New York Times, 01/03/1995) confirme nos résultats "although smoking interferes with weight gain, it does not shorten pregnancy." Concernant la relation entre niveau d'étude et tabagisme, de nombreuses publications officielles [1] font état "de la diminution du tabagisme chez les femmes de niveau social plus élevé [...] les femmes cadres supérieurs fumaient moins que les catégories intermédiaires qui, elles-mêmes, fumaient moins que les ouvrières", reste à prouver que le niveau d'étude et le niveau social sont corrélés...

## 1.2 Données crabs

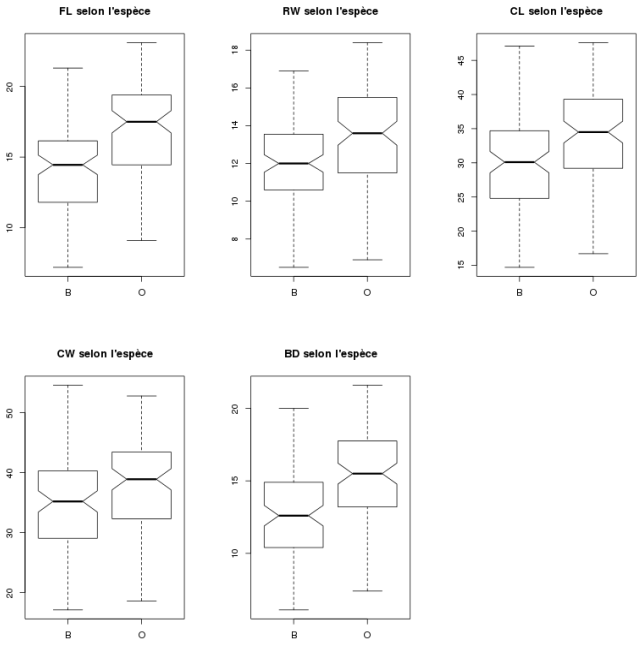
Nous étudions maintenant un jeu de données constitué de 5 mesures morphologiques sur 200 crabes. Nous connaissons l'espèce et le sexe de chacun d'entre-eux. Il y a donc 4 populations (divisées selon le sexe et l'espèce) de 50 individus.

**Aperçu des données**
Avant toute analyse, nous commençons par avoir un aperçu (sur l'ensemble de notre jeu de données) des valeurs prises par les données quantitatives dont nous disposons sur l'ensemble de notre jeu de données. En traçant les boxplot sur les données brutes, nous remarquons la présence d'un outlier dans le cas de RW. C'est le crabe de l'enregistrement 200. Nous décidons de le retirer pour ne pas fausser les analyses à venir. Ensuite, une fois les données centrées, pour mieux comparer la dispersion, nous traçons les boxplot suivants :




Nous remarquons deux groupes parmi les variables, chacun ayant globalement une dispersion similaire de leurs valeurs. FL, RW et BD d'un côté, CL et CW de l'autre. La dispersion des valeurs de CL et CW est plus importante.

**Différences morphologiques selon l'espèce ?**
Afin de déterminer s'il existe des différences de caractéristiques morphologiques selon la variable "espèce", nous comparons d'abord visuellement les distributions des valeurs de chaque variables, un crabe est soit de l'espèce bleue (B) soit de l'espèce orange (O).




Si la dispersion est similaire, les médianes sont différentes. Sauf peut-être pour CW, au vu du chevauchement possible des intervalles de confiance.

Les différences en valeur absolue entre les bornes des intervalles de confiance d'une espèce par rapport à l'autre sont :

| FL | RW | CL | CW | BD |
|---|---|---|---|---|
| 1.58 | 0.499 | 1.23 | 0.152 | 1.47 |

Les médianes sont donc différentes pour toutes les variables, selon l'espèce. **Les crabes bleus ont des valeurs plus élevées pour toutes les variables. En particulier, la taille de leur lobe frontal (FL) est plus importante.**

**Égalité des moyennes** Afin de tester l'égalité des moyennes, pour ensuite confirmer notre hypothèse d'écart des variables selon l'espèce, nous devons d'abord tester la normalité des échantillons. Grâce au test de Shapiro-Wilk, nous obtenons les valeurs de *p-value* suivantes :

| | FL | RW | CL | CW | BD |
|---|---|---|---|---|---|
| bleu | 0.499 | 0.588 | 0.674 | 0.714 | 0.294 |
| orange | 0.156 | 0.352 | 0.604 | 0.409 | 0.469 |

L'échantillon est normal si la *p-value* est strictement supérieure à 0.1. C'est le cas pour toutes les variables. L'hypothèse de normalité est donc acceptée et il est possible d'appliquer le test de Student.

Nous réalisons un test de Student bilatéral sans faire d'hypothèse sur l'égalité des variances (variante dite test de Welch). Le niveau de signification est de 5%.

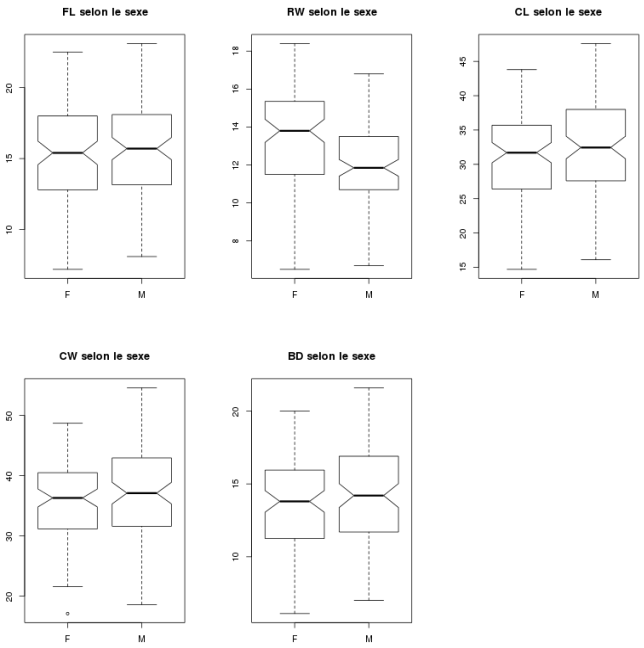
1. notamment celle disponible sur http ://www.ipubli.inserm.fr/bitstream/handle/10608/149/ ?sequence=6

| FL | RW | CL | CW | BD |
|---|---|---|---|---|
| 1.68e-10 | 9.43e-06 | 5.49e-05 | 3.10e-03 | 7.42e-10 |

TABLE 6 – Valeurs des *p-value* du test de Student

Toutes les *p-value* sont inférieures à 0.05. Les crabes des deux espèces ont donc des caractéristiques morphologiques différentes aussi en moyenne.

**Différences morphologiques selon le sexe ?**
Nous procédons de la même manière que précédemment pour la comparaison par rapport au sexe.




Dans ce cas, **seule la largeur de l'arrière (RW) diffère de manière significative**, au sens de la médiane. **Les femelles (F), ont donc une valeur de RW plus élevée**.

Le test de Shapiro-Wilk nous indique que nous pouvons appliquer le test de Student. En effet, aucune *p-value* n'est plus petite que 0.1.

Le test de Student, avec les mêmes paramètres donne les résultats suivants :

| | *p-value* | IC inf | IC sup |
|---|---|---|---|
| FL | 0.441 | -1.35 | 0.590 |
| RW | 5.05e-05 | 0.750 | 2.11 |
| CL | 0.101 | -3.61 | 0.325 |
| CW | 0.228 | -3.52 | 0.842 |
| BD | 0.154 | -1.63 | 0.259 |

Seule la p-value de comparaison sur RW est inférieure à 0.05. Les caractéristiques morphologiques autres que RW ne différent donc pas en moyenne. Cela confirme les résultats obtenus pour la médiane.

Maintenant que nous avons trouvé les différences morphologiques, nous pouvons chercher si à partir de ces informations nous pouvons identifier l'espèce ou le sexe d'un crabe à partir d'une ou plusieurs de ses caractéristiques.

Les différences entre espèces sont les plus élevées pour FL et BD, et les plus faibles pour CW. Comme le montre le garphique matriciel suivant (réalisé avec GGobi), le tracé de FL en fonction de CW (ou l'inverse) montre clairement la séparation, linéaire ici, entre les deux groupes. Chaque groupe étant constitué des individus d'une seule espèce (ici, les crabes bleus sont représentés en jaune).

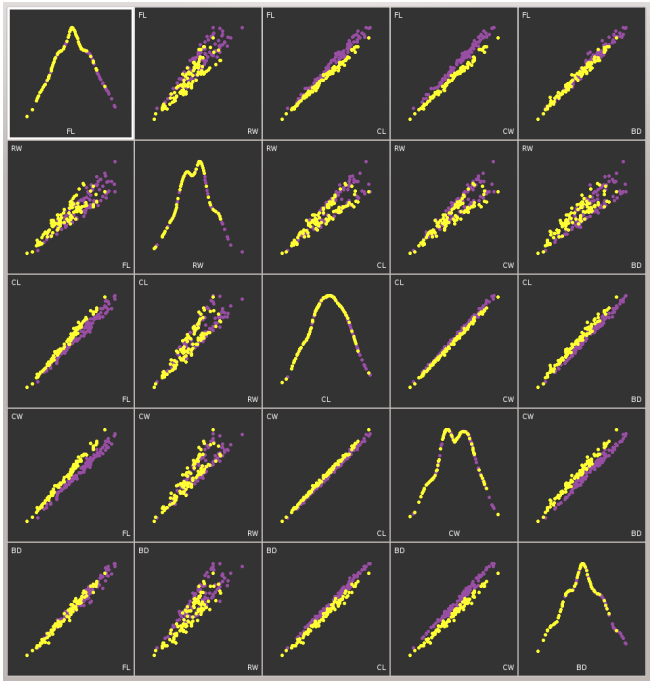
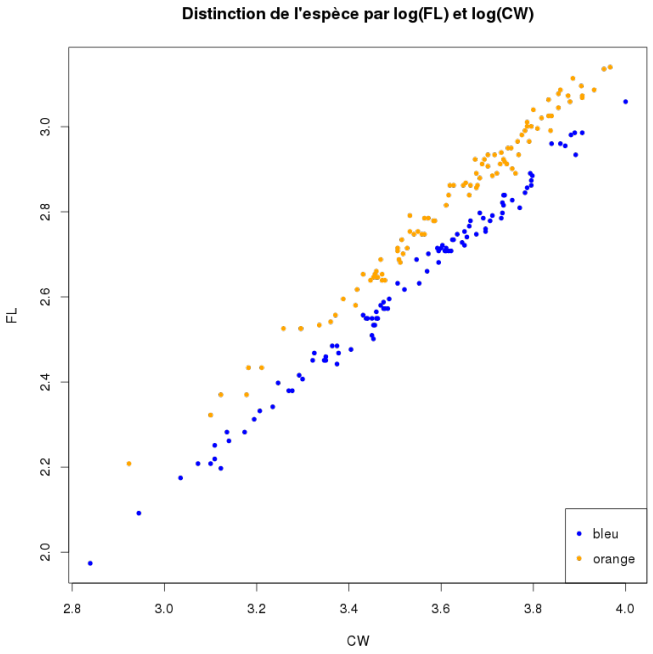



FIGURE 1 – scatterplot matrix avec brushing (espèce bleu en jaune)

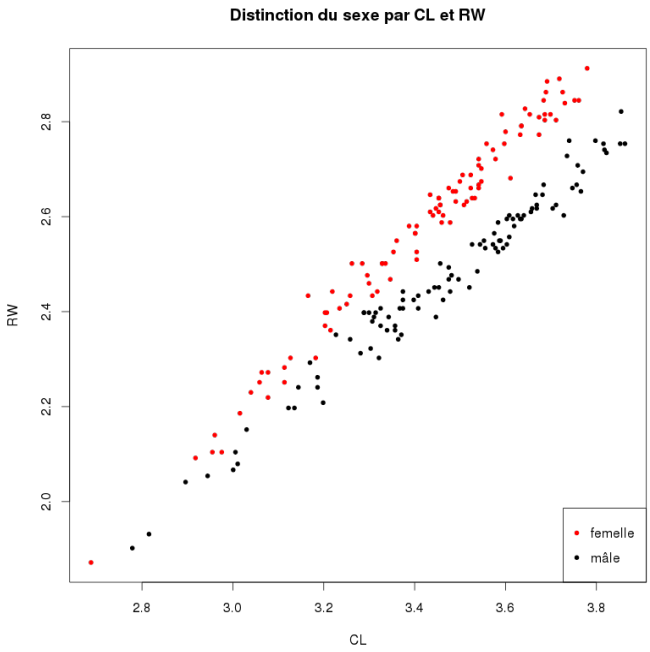
Utiliser BD à la place de FL montre aussi une bonne séparation. Cependant, le tracé individuel de FL en fonction de CW et BD en fonction de CW, montre que dans le cas de BD, la séparation n'est pas parfaite contrairement à l'autre cas. Pour exprimer plus explicitement la corrélation linéaire qui existe entre FL et CW, nous passons au logarithme.




Si nous traçons une droite entre ces deux nuages de points, le placement par rapport à cette droite d'un nouveau crabe nous permettra de dire si c'est un crabe de l'espèce bleu ou orange.

Pour la distinction selon le sexe, seul RW est significativement différent. Les deux autres caractéristiques

à montrer le plus de différence selon le sexe, mais sans que cela soit significatif, sont CL et CW. Le tracé de CL ou CW en fonction de RW fait apparaître une séparation, mais elle devient nette seulement à partir d'un certain seuil, comme nous pouvons le voir sur la figure ci-dessous.




L'exactitude de la détermination du sexe selon la valeur de CL et RW sera donc moins précise pour de petites valeurs.

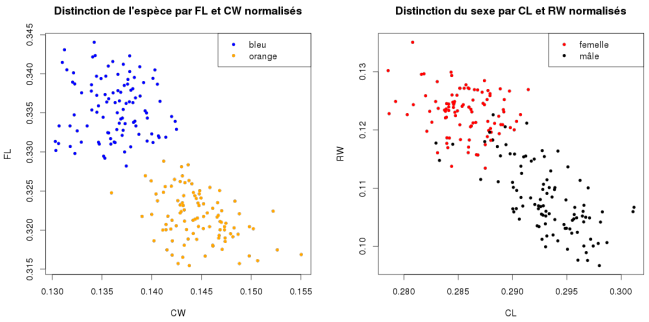
La présence de nuages de points presque assimilables à une droite sur le graphique matriciel montre une très forte corrélation entre les variables. Ceci est confirmé par le calcul des coefficient de corrélation, tous supérieurs à 0.887.

| | FL | RW | CL | CW | BD |
|---|---|---|---|---|---|
| FL | 1.000 | 0.905 | 0.978 | 0.964 | 0.987 |
| RW | 0.905 | 1.000 | 0.892 | 0.899 | 0.887 |
| CL | 0.978 | 0.892 | 1.000 | 0.995 | 0.983 |
| CW | 0.964 | 0.899 | 0.995 | 1.000 | 0.967 |
| BD | 0.987 | 0.887 | 0.983 | 0.967 | 1.000 |

TABLE 7 – Coefficients de corrélation entre les variables

Les couples de variables choisis pour l'identification de l'espèce ou du sexe semble être ceux de corrélation la plus basse par rapport aux autres, sauf pour RW avec BD, CW ou FL.

La très forte corrélation vient du fait que les proportions du corps sont habituellement respectées dans la nature. Une manière de s'affranchir de ce phénomène est de diviser chaque valeur par la somme de toutes celles de l'individus. Diviser par BD ou CW permet aussi de décorréler les variables. Ceci se voit très rapidement sur des graphiques matriciels. Nous pouvons donc supposer que la taille du corps (BD) et la largeur de la carapace (CW) sont des caractéristiques auxquelles les autres sont proportionnelles. => proportionnelles ? Après traitement, les graphiques de disctinction précédents donnent ceci :




Nous retrouvons la séparation parfaite pour la disctinction de l'espèce et imparfaite pour le sexe. Une fois les données réparties dans l'espace de cette manière, il sera possible d'appliquer des méthodes de type k-means.

# 2 Analyse en composantes principales

Le but de cette partie est d'appréhender l'analyse en composantes principales, qui permet de traiter des données multidimensiennelles d'un espace très large de variable en réduisant cet espace, en conservant au mieux l'information.

## 2.1 Exercice théorique

Les données sont contenues dans un tableau de 4 individus 3 variables.

Afin d'obtenir les axes factoriels, on commence par centrer la matrice donnée, on peut ainsi calculer la matrice de variance S = $\frac{1}{n}$ .Y'.Y = $\frac{1}{4}$ .Y'.Y =

$$\begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 1.5 & -0.5 \\ 0 & -0.5 & 1.5 \end{pmatrix}$$

Suite à la diagonalisation de la matrice, nous obtenons les valeurs propres et axes suivants :

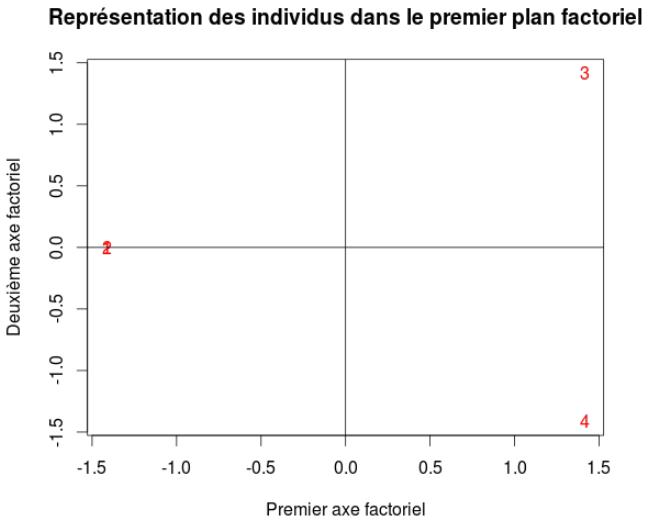
| | C1 | C2 | C3 |
|---|---|---|---|
| valeurs propres | 2.0 | 1.0 | 0.5 |
| % d'inertie | 57.14 | 28.57 | 14.29 |
| % d'inertie cumulés | 57.14 | 85.71 | 100.00 |

Les deux premières dimensions aggrègent presque 86% de l'information, on peut donc représenter 86% de l'information sur le premier plan factoriel défini par les deux premiers axes.

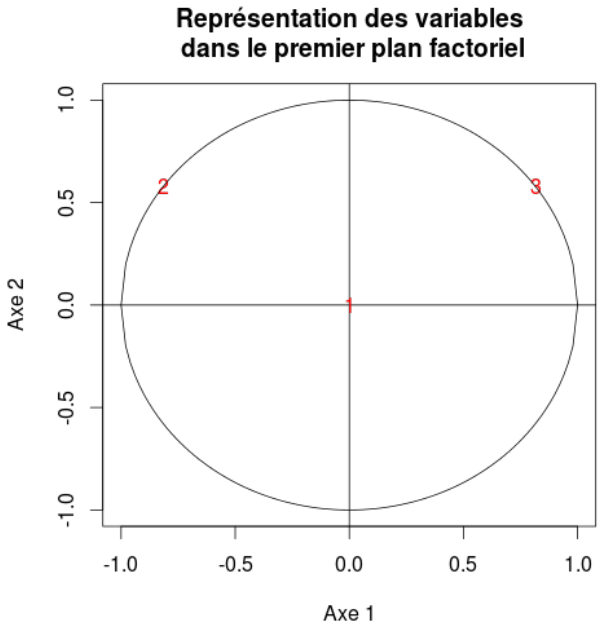
On calcule les composantes principales :

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

La représentation des quatres individus dans le premier plan factoriel nous montre que les deux premiers individus sont représentés par le même point. En effet, la variable qui les différencie n'est pas représentée par le premier plan factoriel : seul le troisième axe la représente.




Grâce à la représentation des variables dans le premier plan nous pouvons effectivement observer que la variable 1 (à l'origine du repère) n'est pas du tout représentée, par aucun des deux premiers axes. On note aussi que les variables 2 et 3 sont très peu voire pas corrélées selon ces axes (elles forment un angle droit). En revanche on ne peut rien affirmer quand à la corrélation entre 1 et 3 et entre 1 et 2 pour la raison citée précédemment de la non représentation de 1.




Grâce à la formule de reconstitution, la matrice reconstituée selon les deux premiers axes obtenue est la suivante :

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

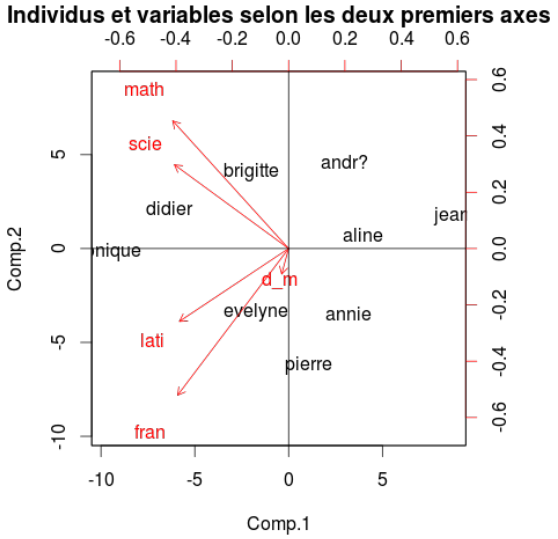
Ces données sont donc reconstituées à 86 %.

## 2.2 Utilisation des outils R

La fonction *princomp* permet d'obtenir les valeurs propres en mettant au carré sa valeur de retour *sdev* (l'écart-type). De la même manière sa valeur de retour *loadings* constitue les axes principaux. La fonction *summary.princomp* donne l'inertie expliquée et les composantes pricipales. La valeur de retour *scores* de *princomp* donne la matrice des composantes principales.

La fonction *plot* appliquée sur le résultat de *princomp* affiche les valeurs propres (variances) associées à chaque composante (une barre par valeur de valeur propre). Cette visualisation donne déjà un apperçu des composantes qui pourront être gardées pour la représentation finale. *biplot* permet de représenter simultanément des individus et des variables (colorées en rouge par défaut). La fonction *biplot.princomp* donne accès à des options supplémentaires et notamment l'option *scale* qui permet d'obtenir une représentation standard.

Dans le cas des données étudiées, on peut observer sur le biplot obtenu que l'axe 1 représente le caractère de "bon élève" (bon dans toutes les matières, dessin-musique excepté) de l'élève, l'axe 2 représente ses capacités en matières scientifiques ou littéraires (sciences et maths dans une direction, latin et français dans l'autre).



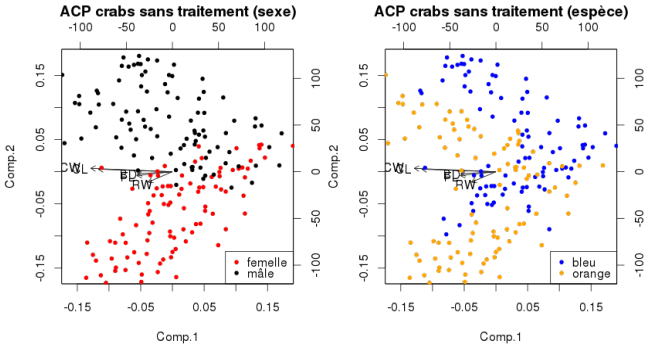

## 2.3 Traitement des données crabs

Nous constatons que le premier axe factoriel représente presque toutes les variables : 98.2% de l'information est expliquée.

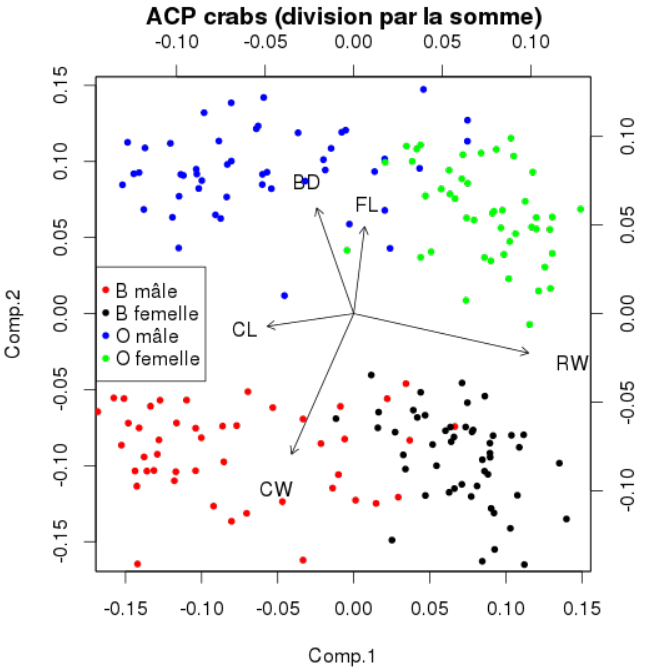
La représentation graphique, donnée par la fonction R *biplot*, indique que CW et CL sont les plus représentées sur cet axe, viennent ensuite FL et BD. RW est par contre un peu représenté par le second axe factoriel.

Ainsi, dans le cas des femelles, qui possèdent une valeur de RW plus élevées (voir conclusion 1.2), cette représentation permet de les situer dans la partie d'ordonnée négative.

Dans le cas des espèces, les valeurs de distinctions trouvées (FL/CW ou BD/CW) sont représentées presque sur le même axe factoriel. Cependant, un léger angle existe entre ces deux couples de valeurs. C'est ce qui donne cette forme en boomerang. L'impossibilité de déterminer l'espèce pour certaines valeurs se retrouve pour les points d'ordonnée proche de 0 (l'angle est nul).




La réduction des données donne une meilleure distinction visuelle dans le cas du sexe mais pas des espèces. Pour avoir une amélioration globale, il faut chercher à expliquer la majorité de l'information sur deux axes et que chaque axe explique environ la même quantité d'information. Par exemple, en reprenant nos conclusions précédentes, si l'on divise l'ensemble des variables par CW, les deux premiers axes expliquent 0.7 point de plus d'information. Enfin, la différence d'information expliquée par chaque axe est 5.14 fois plus importante lorsque nous divison l'ensemble des variables par la somme des valeurs. Nous retenons donc cette solution dernière et obtenons ainsi une représentation visuelle claire, qui permet la distinction entre les espèces et le sexe.




FL et CW rendent bien la distinction des espèces possibles et CL et RW celle du sexe. Nos résultats sont cohérents avec 1.2.