

Premier Rendu

Oumaima Talouka, Zineb Slam

2 avril 2017

Dans ce rapport du TP1 de l'UV SY09 nous allons expliquer notre démarche dans l'analyse des données en expliquant les résultats obtenus. Ce TP est composé de 2 parties. La première partie a pour objectif de se familiariser avec les méthodes de traitement et de visualisation de données sur R. La deuxième partie traite de l'Analyse en composantes principales (ACP). Nous allons travailler avec 3 dataset notes de SY02, Crabs et Pima qu'on va d'abord analyser et décrire avant d'y réaliser l'ACP en seconde partie. Pour les graphes obtenus nous avons utilisés la librairie ggplot2 qui offre un grand nombre de fonctionnalité Le code R sera fourni en annexe.

1 Statistique descriptive

1.1 Notes SY02

Le dataset "*sy02-p2016*" représente les notes des étudiants de l'UTC en méthodes statistiques durant le printemps 2016. Nos données comptent $N= 296$ individus (étudiants) et $p= 11$ variables. Nous avons omis les étudiants qui n'avaient pas de résultats (NA) ou qui ont 'ABS' en résultat de l'UV parce que ce qu'on a pense que ce sont des étudiants qui s'étaient de-inscrits ou n'ont pas suivi l'UV et donc ne constituent pas d'information pour notre dataset. Nous gardons néanmoins les étudiants qui ont NA en médian ou en final mais qui ont obtenu un résultat final.

```
1 dataset = dataset[!is.na(dataset$resultat), ]  
2 dataset = dataset[dataset$resultat != 'ABS', ]
```

Nous gardons donc $N = 284$ individus pour notre étude

1.1.1 Description des variables

- **Variables Quantitatives** : *note.median*, *note.final* et *note.totale*.
- **Variables Qualitatives Nominale** : *nom*, *specialite*, *status*, *dernier.diplome.obtenu* et *correcteur.median*, *correcteur.final*
- **Variables Qualitatives Ordinales** : *niveau* et *resultat*

Nous allons à présent définir chaque variable en précisant son intervalle ou les valeurs qu'elle peut prendre.

- **Nom** : chaîne de caractère identifiant chaque étudiant.
- **Spécialité** : la branche de l'étudiant : *GB*, *GM*, *GSM*, *GP* et *GI*.
- **Niveau** : Semestre d'étude de chaque étudiant de 1 à 6.
- **Statut** : Soit l'étudiant est de l'*UTC* ou en *semestre d'échange*
- **dernier.diplome.obtenu** : *BAC*, *DUT*, *CPGE*, *ETRANGER SUPERIEUR*, *LICENCE*, *OTHER*, *NA*
- **Note Médian** : note de l'examen Médian, ensemble de réel de 0 à 20.
- **Correcteur Médian** : ID du correcteur du médian {*Corr1*, *Corr2*, *Corr4*., *Corr5*,...,*Corr8*}
- **Note.final** : Note de l'examen final, ensemble de réel de 0 à 20.
- **Note.totale** : Note totale obtenue à partir de la note du médian et la note du final
- **Correcteur.final** : identifiant du correcteur du final. {*Corr1*..*Corr3*, *Corr4*., *Corr5*,...,*Corr8*}
- **Résultat** : Résultat obtenu en SY02 : *A*, *B*, *C*, *D*, *E* et *FX* et *FX.ABS*. *ABS* est pour indiquer que l'étudiant était absent pour l'examen en question.

Pour les notes de médian et de final il y'a des notes non mentionnées (*NA*) par contre tous les étudiants ont un résultat final c'est pour cela on n'a pas enlevé les étudiants avec *NA* en médian ou/et en final.

Les variables importantes dans ce dataset sont les résultats des étudiants et comment ceux-ci sont influencés par d'autres variables, par exemple le niveau et la spécialité

Il est évident qu'il existe une relation linéaire entre ces 3 variables : *note.median*, *note.final* et *note.totale* vu que la note totale est exprimée par une relation linéaire entre la *note.median* et la *note.final* (par exemple $note.totale = 40\% * note.median + 40\% * note.final + cste$). La variable *note.totale* et la variable *resultat* sont des variables fortement corrélées. En effet le résultat est une "traduction" de la note totale. On pourrait éventuellement se demander sur la relation entre les notes et le niveau, la spécialité, le diplôme ainsi que le correcteur. C'est ce qu'on va essayer d'analyser dans ce qui suit.

1.1.2 Liens entre les variables

Après avoir représenté la matrice de graphes pour chaque variable nous avons pu observer que les variables : *note.median*, *note.final*, et *note.totale* sont linéaires ce qui confirme notre hypothèse précédente. Le graphe matriciel ci-dessous

présente aussi les coefficients de corrélation entre chacune de ces variables.

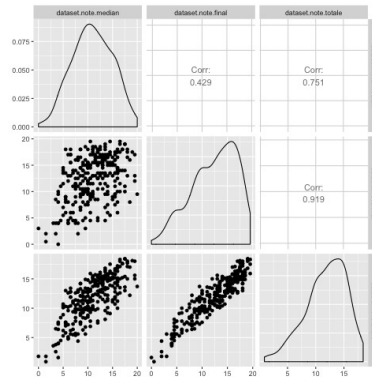


FIGURE 1 – Plot general

1.1.3 Homogénéité et Distribution des notes

La figure ci-dessous représente trois diagrammes a boites des notes de médian, final et le résultat de l'UV SY02.

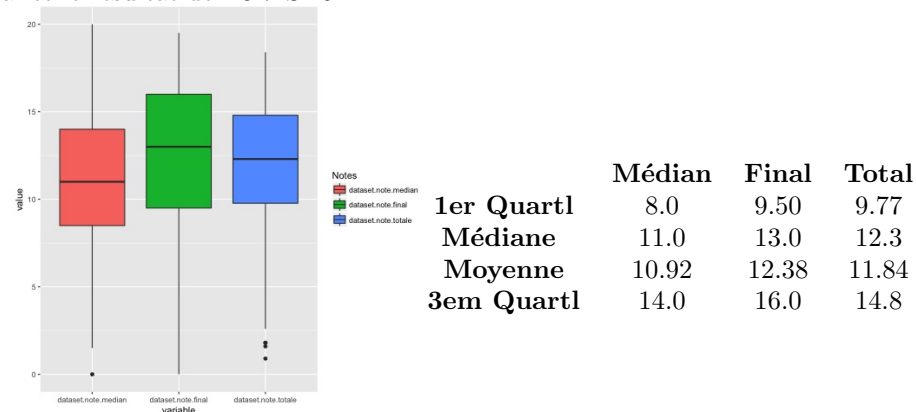


FIGURE 2 – Boxplot des Notes

On remarque les boites sont petites ce qui laisse penser que la variance des résultats est relativement petite et donc aussi que le niveau des étudiants en statistiques l'est aussi. En comparant les diagrammes du final et du médian on remarque que les notes ont augmenté. Enfin si on analyse le dernier diagramme on voit que les deux boites du part et d'autre de la médiane ont la même taille, on peut donc mettre l'hypothèse que les résultats sont normalement distribués.

Il est tout à fait logique que le diagramme de boites des notes finales se situe entre les 2 puisque que la note finale est une moyenne des deux notes du médian et du final.

1.1.4 Lien entre la réussite, la formation, la branche et le niveau

Comme les étudiants de chaque branche n'ont pas les *même effectifs* on a choisi de représenter les données sous forme de diagrammes de moustache pour mieux pouvoir les comparer.

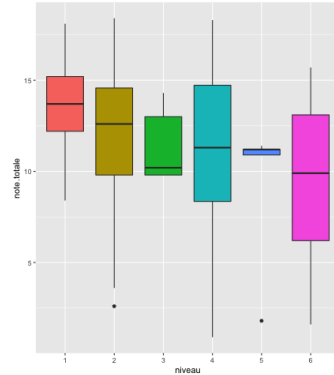


FIGURE 3 – Diagramme en boîte de lien entre le niveau et le resultat

D'après la figure 3 on remarque que les étudiants venant durant les premiers semestres sont ceux qui réussissent le mieux leur examens de SY02, suivi par ceux en GX02. On remarque que les notes des étudiants en GX04 et GX05 ont une grande variance. Les étudiants en GX04 et en GX05 ont . Il faut aussi remarquer que les étudiants en 4em et 5 em sont peu comme c'est des semestres de départ en stage.

En ce qui concerne l'influence de la spécialité sur les notes on remarque qu'il y'a une grande variance dans les notes chez les étudiants en TC et ISS et GSM, contrairement au GI et GP qui ont plutôt un niveau homogène et qui semble bien réussir l'UV. De plus les TC sont aussi ceux qui réussissent le mieux l'UV.

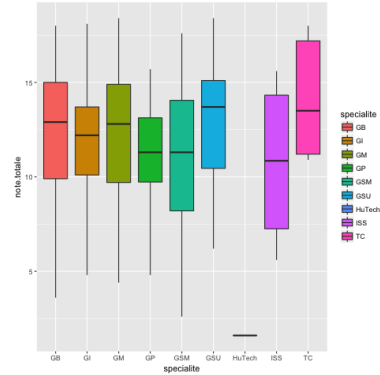


FIGURE 4 – Diagramme en boîte de tre la branche et le resultat

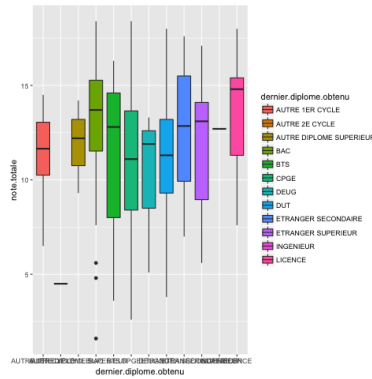


FIGURE 5 – Diagramme de moustache de lien entre la formation et le resultat

On remarque qu'il y'a une grande variance chez presque tous les diplômes

obtenus. Les élèves qui réussissent le mieux et le plus sont ceux provenant de la Licence. Plus de 75% (**Check**) des étudiants de autre premier cycle et autres diplômes et du BAC réussissent l'UV SY02. En ce qui concerne les résultats des étudiants pour le reste des diplômes ils ont une grande variance et ne sont pas normalement distribués. En particulier les étudiants du BTS, CPGGE et du DEUG sont ceux qui ont la plus grande variance et le plus grand échec.

1.1.5 Influence du correcteur sur la note

Les deux diagrammes ci-dessous montrent la dispersion des notes de final et de médian pour chaque correcteur.

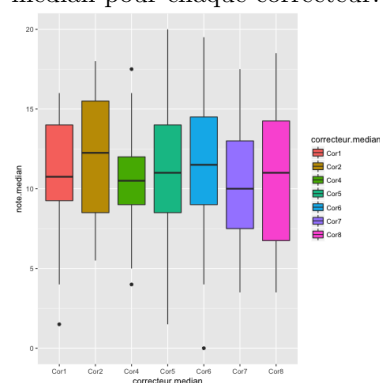


FIGURE 6 – Diagramme de moustache des notes de median en fonction des correcteurs

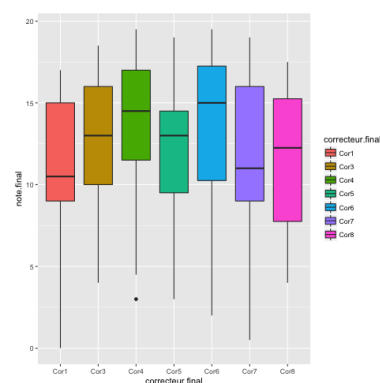


FIGURE 7 – Diagramme de moustache des notes du final en fonction des correcteurs

on considère que les copies sont aléatoirement distribuées pour chaque correcteur et donc il n'y a pas de correcteur qui a que les 'bons' ou les 'mauvais' étudiants. À première vue on remarque qu'en général les notes sont dispersées pour chaque correcteur et donc il n'y a pas vraiment de correcteur en particulier qui semble être plus 'sévère' que les autres. On peut émettre l'hypothèse que le *Corr4* a été plus stricte dans ses corrections de médians, néanmoins son diagramme de notes de final prouve le contraire. Il se peut donc que le correcteur avait les copies des 'mauvais' étudiants par fruit de hasard. Finalement, comme il a été cité précédemment les notes ont augmenté au final par rapport au médian, on observe donc bien que ceci est le cas pour tous les correcteurs.

1.1.6 Conclusion

Cette première analyse data nous a permis d'étudier quels sont les facteurs qui influent sur la réussite d'un étudiant dans l'UV SY02 comme le dernier diplôme obtenu ou le niveau et ceux qui n'influencent pas comme le correcteur. Néanmoins ces conclusions sont propres à la population du P16 et donc biaisées; pour pouvoir généraliser il faut analyser les notes de SY02 sur plusieurs semestres avec des populations différentes.

2 Crabs

Le dataset "*Crabs*" représente un jeu de données de 200 crabes décrits par huit variables, trois sont qualitatives et cinq sont quantitatives.

2.1 Description des variables

- **Variables Qualitatives Nominales** : crabs.sp, crabs.sex, crabs.inde
- **Variables Quantitatives** : crabs.FL, crabs.RW, crabs.CL, crabs.CW, crabs.BD

Nous pouvons représenter les données des variables quantitatives à l'aide d'un boxplot.

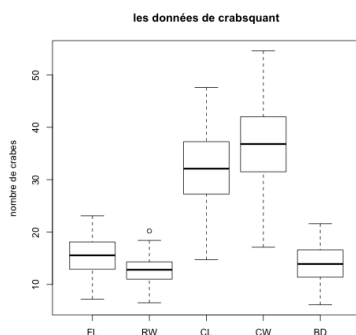


FIGURE 8 – Boxplot des données quantitatives

Nous remarquons d'ores et déjà que deux catégories de variables se distinguent, d'un côté FL, RW et BD et d'un autre, CL et CW.

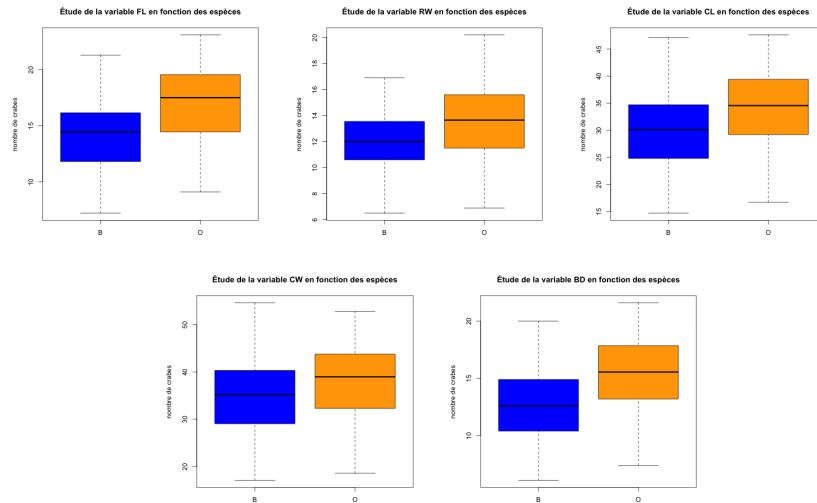
Avant de continuer l'analyse de ces données, nous pouvons préciser la signification de chacune de ces variables comme suit :

- **sp** : (*species*), espèce, "B" pour Bleu et "O" pour Orange
- **sex** : sexe, "F" pour Féminin et "M" pour masculin
- **index** : index de 1 à 50 pour chacune des 4 catégories suivantes : {"B,M", "O,M", "B,F", "O,F"}
- **FL** : Frontal Lobe size en mm
- **RW** : Rear Width en mm
- **CL** : Carapace Length en mm
- **CW** : Carapace Width en mm
- **BD** : Body Depth en mm

2.2 Analyse descriptive des données

2.2.1 Représentation de chaque caractéristique en fonction de l'espèce

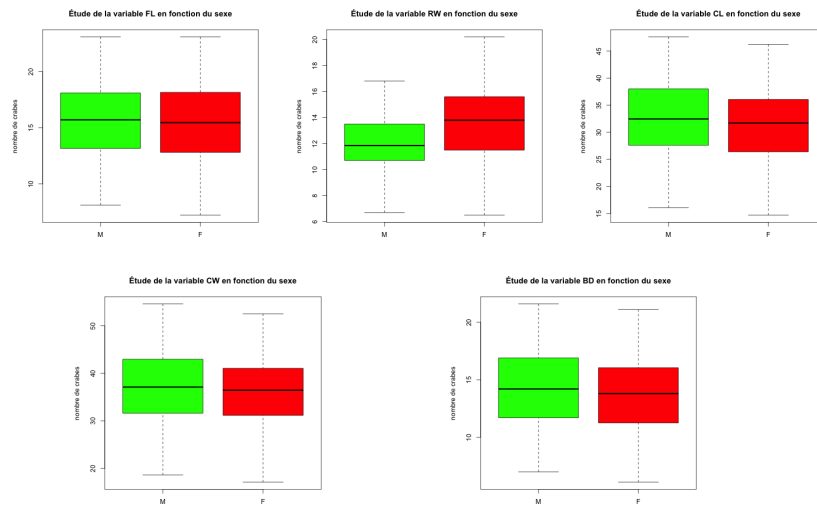
Afin de voir s'il y a une différence de caractéristiques morphologiques en fonction de l'espèce d'abord, nous représentons les boîtes à moustache de chaque variable morphologique en fonction de la variable sp comme suit :



Il y a des distributions assez différentes en fonction de l'espèce. Les intervalles de confiance ne se chevauchent pas. Cependant, la dispersion des données reste assez homogène au vu des boxplots.

2.2.2 Représentation de chaque caractéristique en fonction du sexe

Afin de voir s'il y a une différence de caractéristiques morphologiques en fonction du sexe, nous représentons les boîtes à moustache de chaque variable morphologique en fonction de la variable sexe comme suit :



A l'opposé des boxplots en fonction de l'espèce, ceux en fonction du sexe relèvent une similitude de la distribution des différentes caractéristiques pour la

plupart, ainsi que la dispersion des données. Nous observons que la variable RW se distingue des autres avec une largeur de l'arrière importante chez les femmes que chez les hommes.

Enfin, la nature de l'espèce impacte les caractéristiques morphologiques, à la différence du sexe, qui lui n'influe que peu ces caractéristiques.

2.2.3 Lien entre les variables

Nous pouvons représenter chacune de ces variables quantitatives en fonction de l'espèce puis en fonction du sexe afin de déterminer la possibilité d'identifier l'une ou l'autre à partir des caractéristiques morphologiques.

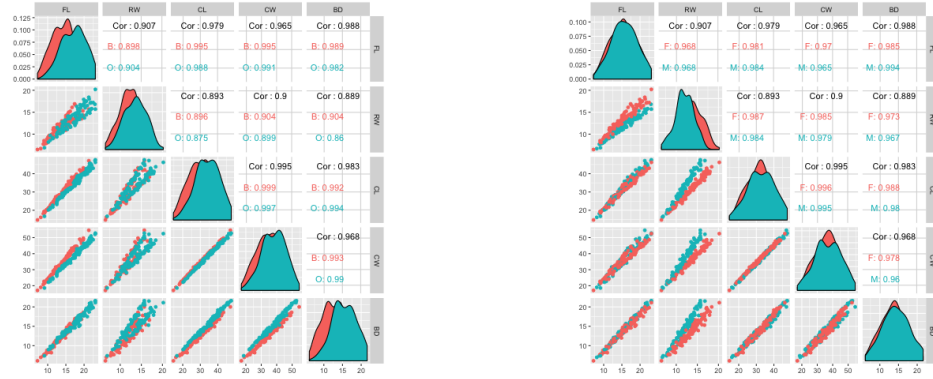


FIGURE 9 – Plot general en fonction de l'espèce (gauche) et du sexe (droite)

Nous remarquons que ni l'espèce ni le sexe ne peuvent vraiment être identifiés à partir d'une ou de plusieurs caractéristiques morphologiques. En effet, dans les cas, l'ensemble des points est représenté sur une même droite, on ne peut pas clairement distinguer la différence. De ce fait, il est difficile de reconnaître une espèce selon ses caractéristiques morphologiques.

2.3 Analyse de la corrélation

	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

FIGURE 10 – Corrélation entre les variables

Il y a une forte corrélation positive entre toutes les combinaisons de variables, telle que la valeur minimale observée est 0.889. Il s'agit de la taille des membres du corps d'un crabe, il semble donc logique et naturel qu'elles soient

proportionnelles entre elles. Une des façons pour s'affranchir de ce phénomène est de diviser chaque valeur par la somme totale de toutes celles de l'individu.

3 Pima

Le dataset "*Pima*" représente un jeu de données constitué de 532 individus tous de sexe féminin décrits par huit variables dont une qualitative et sept sont quantitatives.

3.1 Description des variables

- **Variables Qualitatives Ordinale** : Pima.z
- **Variables Quantitatives** : Pima.npreg, Pima.glu, Pima.bp, Pima.skin, Pima.bmi, Pima.ped, Pima.age

Nous pouvons représenter les données des variables quantitatives à l'aide d'un boxplot.

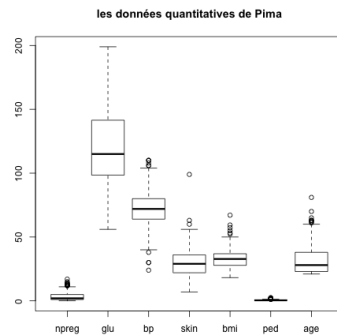
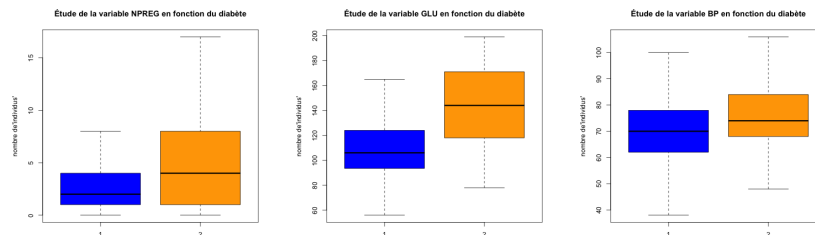
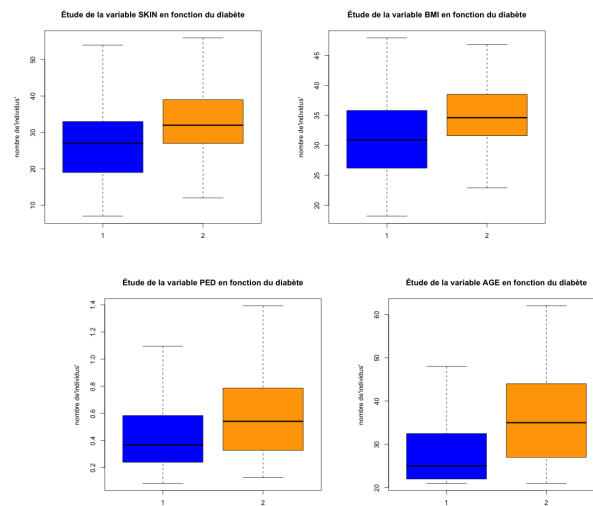


FIGURE 11 – Boxplot des données quantitatives

3.2 Analyse descriptive des données

3.2.1 Représentation de chaque caractéristique en fonction de z (diabétique ou pas)





3.2.2 Lien entre les variables

3.3 Crabs

3.4 Pima

3.5 Crabs

3.6 Pima

4 Analyse Composantes Principales

Dans cette section nous allons nous concentrer sur la technique de l'Analyse en Composantes Principales qui consiste à transformer des variables corrélées en nouvelles variables non corrélées.

4.1 Exercice théorique

Nous travaillons à présent sur les données notes mais cette fois-ci les correcteurs sont les individus et les variables : *moy.median*, *std.median*, *moy.final*, *std.final*. On associe les mêmes pondérations à tous les individus, et on munit R^p de la métrique euclidienne. Dans la suite la matrice M de pondération des variables est la matrice identité

4.1.1 Calcul des axes factoriels de l'ACP

Avant de commencer le calcul on commence par centrer notre matrice de données X.notes en soustrayant la moyenne de chaque colonne ce qui permet de mettre le centre de gravité du nuage de points à l'origine. Ceci est possible grâce à la fonction *scale* de R. Tout d'abord on calcule la matrice des variances M par la formule suivante : $V = \frac{1}{6} * X.notes * X.notes^T$. Nous obtenons ensuite les valeurs propres $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ triées par ordre grâce à la fonction *eigen* ainsi que les vecteurs propres associés. Nous obtenons la matrice U (*ci-dessous*) de vecteurs qui constitue nos axes factoriels de l'ACP.

$$\lambda_1 = 1.52, \lambda_2 = 1.03, \lambda_3 = 0.62, \lambda_4 = 0.15$$

$$U = \begin{bmatrix} u1 & u2 & u3 & u4 \\ -0.5691089 & 0.3831062 & -0.4770578 & 0.54932742 \\ -0.6626009 & -0.2607092 & -0.1954537 & -0.67438013 \\ 0.2342134 & 0.8173384 & -0.1888389 & -0.49136735 \\ -0.4268714 & 0.3423716 & 0.8357952 & 0.04482142 \end{bmatrix}$$

Nous allons à présent calculer le pourcentage cumulé d'inertie expliquée pour chacun des axes u1, u2, u3 et u4. Pour cela on utilise la formule : $P_k = \frac{\sum_{i=1}^k \lambda_k}{\sum_{i=1}^p \lambda_k}$. Nous obtenons alors l'histogramme ci-dessous. Nous remarquons que les 3 premières composantes ensemble expliquent 95%, on pourra donc ne garder que les 3 premières axes.

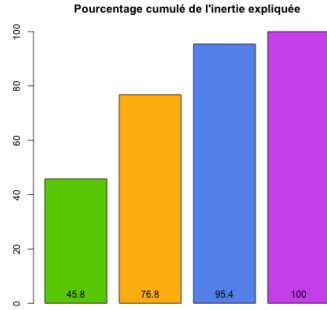


FIGURE 12 – Pourcentage cumulé de l'inertie expliquée

4.1.2 Composantes Principales

Soit C la matrice avec en colonnes les coordonnées de chaque vecteur de composante principale. C s'exprime comme $X.notes.acp = X.notes * U$. Remarquons aussi que la matrice C sera aussi centrée

$$X.notes.acp = \begin{bmatrix} c1 & c2 & c3 & c4 \\ 0.1669338 & -0.8150980 & 0.4438352 & 0.75983420 \\ 2.2340546 & 0.9363064 & 0.1554035 & 0.06367979 \\ -0.1186174 & -0.7571973 & -1.6361255 & -0.05651747 \\ -1.3438493 & 1.8308473 & -0.1804311 & 0.01407302 \\ 0.7303288 & -0.7032237 & 0.4432574 & -0.50642647 \\ -1.3349828 & -0.4916345 & 0.7740606 & -0.27464306 \end{bmatrix}$$

4.1.3 Représentation des Individus et variables dans le plan factoriel

Afin de connaître la contribution de chaque composante dans une variable nous calculons la matrice de corrélation entre l'ancienne matrice des données $X.notes$ et la nouvelle matrice avec les composantes. Ceci se fait grâce à la fonction *corr*. Nous obtenons la matrice de corrélation ci-dessous.

	moy.median	std.median	moy.final	std.final
c1	-0.775	-0.90	0.32	-0.58
c2	0.43	-0.29	0.91	0.38
c3	-0.41	-0.17	-0.16	0.72
c4	0.23	-0.29	-0.21	0.02

On retrouve encore l'un autre argument comme quoi les 3 premières composantes suffisent pour expliquer les données. En effet d'après la matrice de corrélation on observe que les variables sont fortement corrélées aux 3 premières composantes. La variable *moy.final* par exemple est fortement corrélée à la 2nd composante. Ceci veut dire que si la composante *c2* est grande ceci veut dire que la moyenne

du final sera aussi importante. Les 2 figures suivantes sont une représentation de 6 correcteurs dans le premier plan factoriel (les composantes 1 et 2), alors que la figure 14 représente la contribution de chaque composante pour chacune des 4 variables.

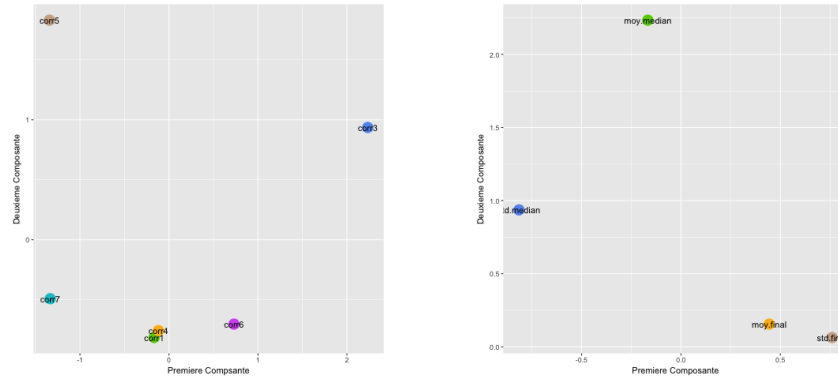


FIGURE 13 – Individus dans le 1er plan FIGURE 14 – Variables dans le 1er plan

4.1.4 Calculer l'expression

4.1.5 Représentation des individus initialement écartés de l'ACP

4.2 Utilisation des outils de R : Crabs

4.2.1 ACP sans traitement préalable

Nous faisons appel à la fonction *princomp* qui calcule les composantes principales de notre dataset. Ensuite nous utilisons le plot dans la figure 18 et le biplot dans la figure 19.

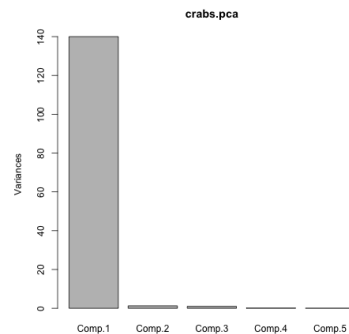


FIGURE 15 – Variance expliquée par les composantes

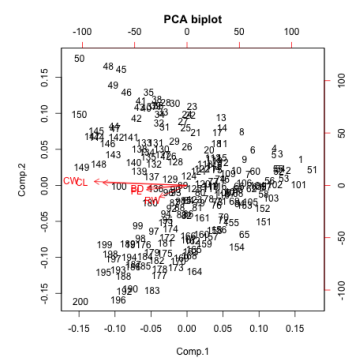


FIGURE 16 – Individus et Variables projetés sur le premier plan factoriel

D'après la figure 18 on remarque que la première composante concentre la majorité de l'information, ceci est aussi illustrée dans la figure 18 puisqu'on

remarque que toutes les variables sont 'encodées' dans la composante 1. D'après notre étude dans la partie analyse des données nous avons remarque que les variables sont fortement corrollees et donc proportionnelles. Or l'une des conditions d'utilisation de l'ACP est que les variables soit decorrelees, ainsi pour de meilleur résultats on essayera dans la question qui suit de decorreler les variables et retenter l'ACP.

4.2.2 Solution proposée

Afin d'obtenir des variables non corolles nous avons divisé chaque donnée d'une ligne par la somme des valeurs de la ligne. De cette manière :

```

1  #nouvelle matrice de donnees avec variables non corolles
2  crabs_decorr = crabs
3  rowsum = rowSums(crabs[,4:8])
4  for(i in 1:n.crabs){
5    crabs_decorr[i,4:8] = crabs_decorr[i,4:8] / rowsum[i]
6  }

```

Ensuite comme nous souhaitons étudier l'espèce et le sexe nous avons pense qu'il serait mieux de regrouper ces 2 resultats en un afin d'obtenir une variable 'réponse' qui prend 4 valeurs ; par exemple la valeur *BF* signifie crabe d'espece B et de sexe Femelle. Les instructions sont détaillées dans le fichier de code en annexe. Le graphe matriciel ci-dessous montres les résultats obtenus.

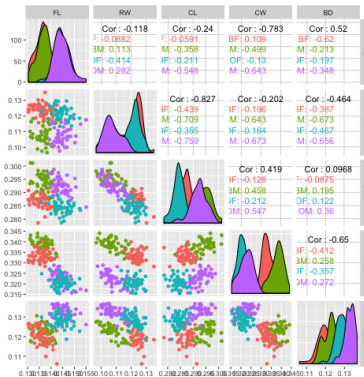


FIGURE 17 – Plot Matriciel des caracteristiques apres decorelation

On remarque qu'a présent on peut visiblement distinguer les espèces par leur caractéristiques Par exemple si on observe le plot de "RW" en fonction de "CW" les espèces sont séparées.

A present nous pouvons réaliser une ACP sur ses données Nous obtenons les résultats suivants :

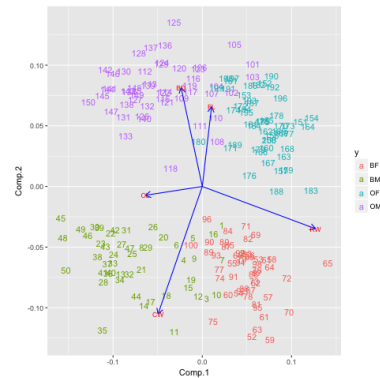
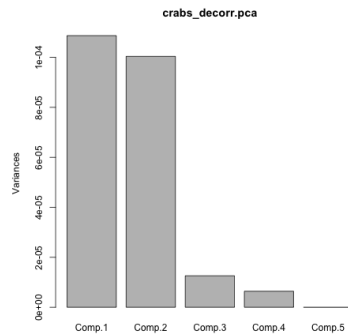


FIGURE 18 – Variance expliquée par les composantes après decorrelation

FIGURE 19 – Individus et Variables sur le 1er plan factoriel apres decorrelation

La figure 19 nous montre que les variables RW et OL sont contenu principalement dans la composante 1 alors que les variables BO, FL et CW seront plus dans la composante 2. Ce qui est aussi affiche dans la 18 vu que les 2 premières composantes ensemble concentre le maximum de variance expliquée Enfin, La même figure nous permet aussi de conclure que **la l'information sur le sexe se trouve dans la composante 1 tandis que la composante 2 contient l'information sur l'espèce du crabe.**

4.3 Pima

5 Conclusion

Ces différents exercices nous ont permis d'abord de manipuler différentes données et de se familiariser avec la notion de variable qualitative et quantitative pour mieux représenter les données graphiquement. Nous avons passé le plus de temps possible dans la première partie afin de mieux comprendre les fonctionnalités du *R* et surtout de *ggplot*, ce qui a facilité notre analyse. Nous avons ensuite utilisé la méthode de l'*ACP* qui est très utile pour l'obtention d'axes decorréles. Enfin si nos variables sont corrélées on peut avoir des valeurs propre nulles et donc un nombre petit de composantes. L'*ACP* est aussi fortement utilisée dans la réduction de dimension, ce qui fera sujet de nos prochaines études en SY09.