

||||| HEAD  
||||| origin/master

=====

UNIVERSITÉ DE TECHNOLOGIE DE  
COMPIEGNE

SY09

DATA MINING

---

## Premier Rendu

---

Oumaima TALOUKA

Zineb SLAM

29 mars 2017



## Résumé

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 utilisé la librairie ggplot2 qui offre un grand nombre de fonctionnalités. Le code R sera fourni en annexe.

# Table des matières

iiiiii HEAD

<b>1</b>	<b>Statistique descriptive</b>	<b>2</b>
1.1	Notes SY02 . . . . .	2
1.1.1	Description des variables . . . . .	2
1.1.2	Analyse descriptive des données . . . . .	3
1.2	Conclusion . . . . .	5
1.3	Crabs . . . . .	5
1.3.1	Description des variables . . . . .	6
1.3.2	Analyse descriptive des données . . . . .	6
1.3.3	Analyse de la corrélation . . . . .	9
1.4	Pima . . . . .	9
1.4.1	Description des variables . . . . .	9
1.4.2	Analyse descriptive des données . . . . .	10
<b>2</b>	<b>Analyse Composantes Principales</b>	<b>11</b>
	=====	
0.1	Notes SY02 . . . . .	2
0.1.1	Description des variables . . . . .	2
0.1.2	Analyse descriptive des données . . . . .	3
0.2	Conclusion . . . . .	7
0.3	Crabs . . . . .	7
0.4	Pima . . . . .	7
<b>1</b>	<b>Analyse Composantes Principales</b>	<b>8</b>
	LLLLLL origin/master	

# Chapitre 1

## Statistique descriptive

### 1.1 Notes SY02

Le dataset "*sy02-p2016*" représente les notes des étudiants de l'UTC en SY02 ( statistiques) durant le printemps 2016. Nos données comptent  $N= 296$  individus (étudiants) et  $p= 11$  variables. **Enlever les étudiants desinscrits**

#### 1.1.1 Description des variables

Variables Quantitatives	note.median, note.final, note.totale
Variables Qualitatives Nominales	nom, specialite, status, dernier.diplome.obtenu, correcteur.median,
Variables Qualitatives Ordinales	niveau, et rltat

TABLE 1.1 – Categorie de Variables

- **Nom** : chaîne de caractère des identifiants de chaque étudiant
- **Spécialité** : la branche de l'étudiant : GB *GtitGSM*, GP, tGP et *GI*.mestre de l'étude de chaque étudiant de 1 à 6.
- **Statut** : Soit l'étudiant est de l'UTC ou en *sd'échange*
- **dernier.diplome.obtenu** : BAC, DUT, CPGE, ETRANGER SUPERIEUR, LICENCE, OTHER, NA
- **Note Médian** : note de l'examen Médian
- **Correcteur Médian** : ID du correcteur du médian de 1 à 8.
- **Note.final** : Note de l'examen final
- **Note.totale** : Note totale obtenue à partir de la note du médian et la note du final
- **Correcteur.final** : D du correcteur du final de 1 à 8.
- **Résultat** : Résultat obtenu de SY02 : A, B, C, D, E et FX

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'autre variable, 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 corrélés. 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 Analyse descriptive des données

#### Lien entre les variables

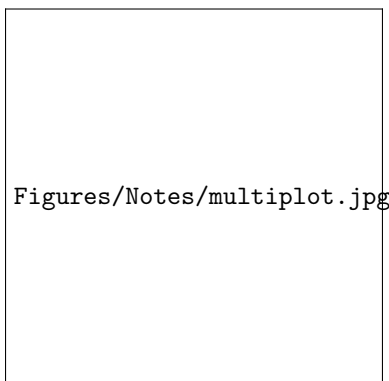


FIGURE 1.1 – Plot general

#### Performance et homogénéité

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

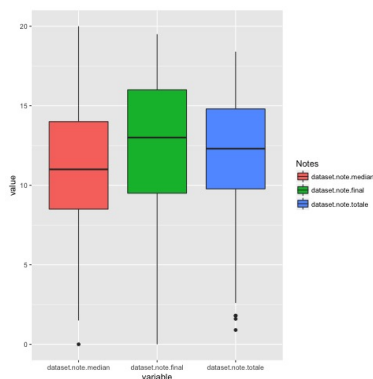


FIGURE 1.2 – Boxplot des Notes de Median, Final et Totale

On remarque que par passage du médian au final les notes ont augmente. En effet ceci est aussi témoigné par le tableau ci-dessous :

Notes	1er Quartl	Medianne	Moyenne	3em Quartl
Median	8.0	11.0	10.92	14.0
Final	9.50	13.0	12.38	16.0
Totale	9.775	12.3	11.845	14.8

TABLE 1.2 – Notes

Il est tout a fait logique que le digramme de résultats se situe entre les 2 puisque que c'est la moyenne **pondérée ??** des 2deux notes du médian etfinal. en entre la réussite, la formation, la branche et le niveau

FIGURE 1.3 – Diagramme a baton de lien entre la formationeet le resultat

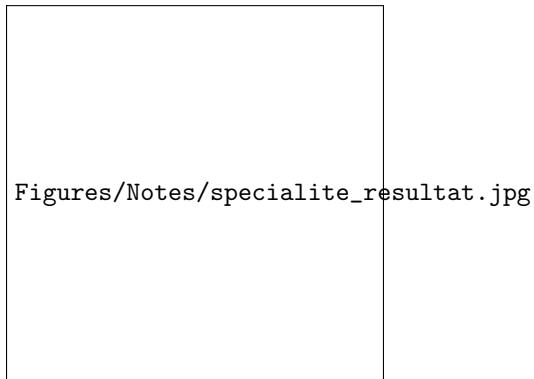


FIGURE 1.4 – Diagramme a baton de lien entre la branche et le resultat

FIGURE 1.5 – Diagramme a baton de lien entre le niveau et le resultat

### Influence du correcteur sur la note

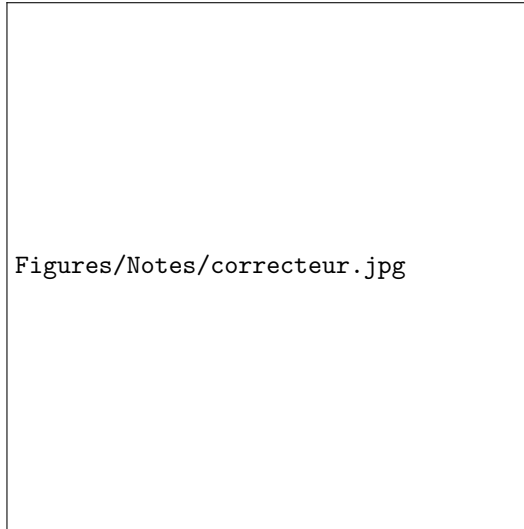


FIGURE 1.6 – Scatterplot des notes en fonction des correcteurs

## 1.2 Conclusion

## 1.3 Crabs

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

### 1.3.1 Description des variables

- **Variables Qualitatives Nominale** : 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.



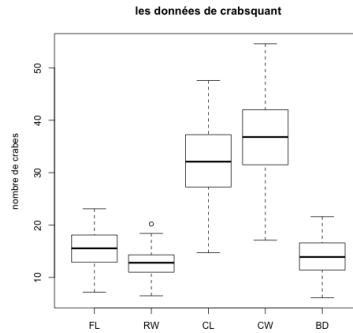


FIGURE 1.7 – Boxplot des données quantitaves

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.

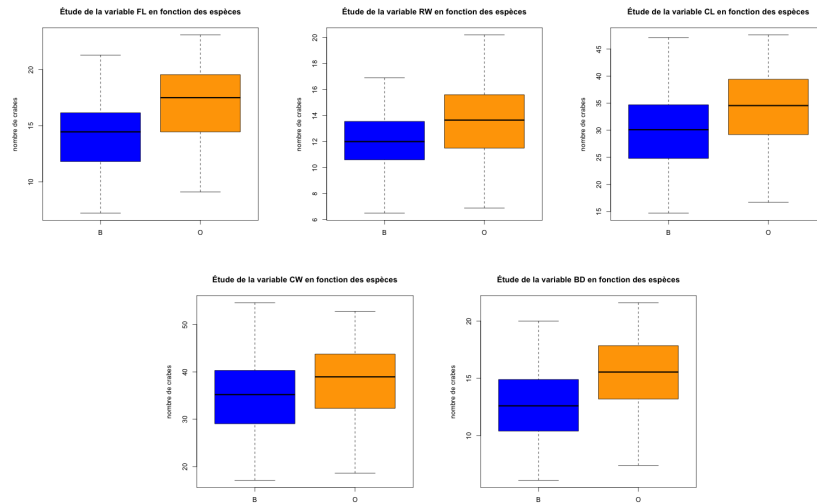
Avant de continuer l'analyse de ces données, nous pouvons préciser la signification de chacune de ces variables comme suit :

- **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

### 1.3.2 Analyse descriptive des données

#### Représentation de chaque caractéristique en fonction de l'espèce

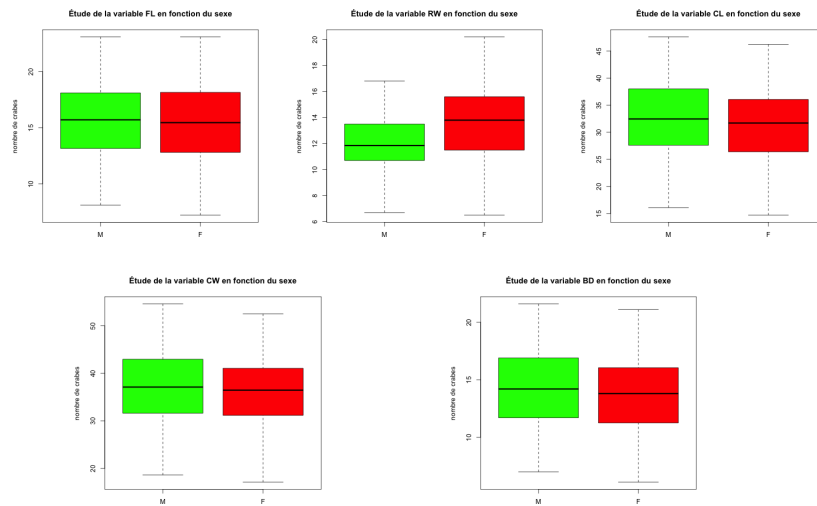
Afin de voir s'il y a une différence de caractéristiques morphologiques en fonction de l'espèce d'abord, nous représentons les boîtes à moustache de chaque variable morphologique en fonction de la variable sp comme suit :



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.

### 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 sex comme suit :



A l'opposé des boxplots en fonction de l'espèce, ceux en fonction du sexe révèlent 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.

### Lien entre les variables

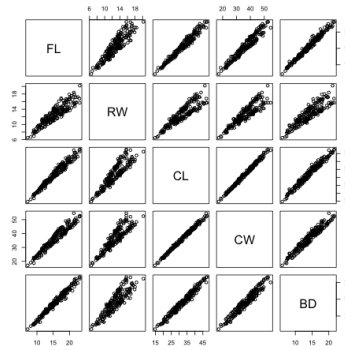


FIGURE 1.8 – Plot general

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.

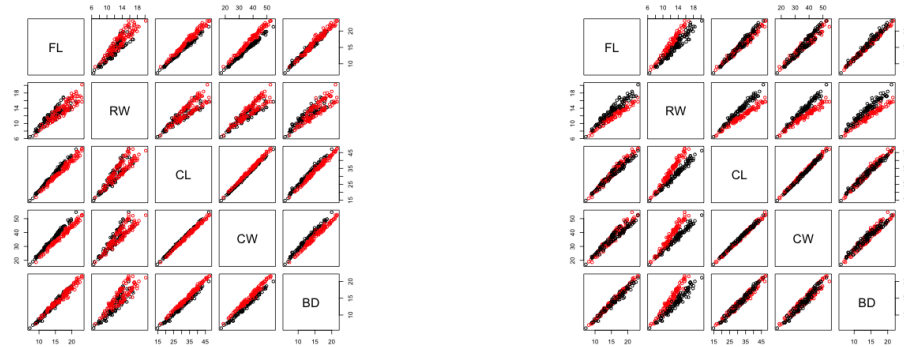


FIGURE 1.9 – Plot general en fonction de l'espèce (gauche) et du sexe (droite)

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 l'une ou l'autre. De ce fait, il est difficile de reconnaître une espèce selon ses caractéristiques morphologiques.

### 1.3.3 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 1.10 – 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.

## 1.4 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.

### 1.4.1 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.

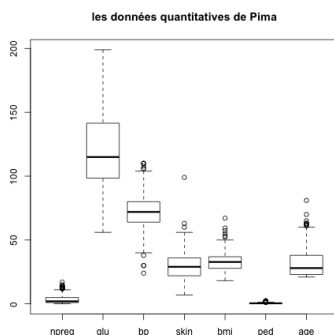
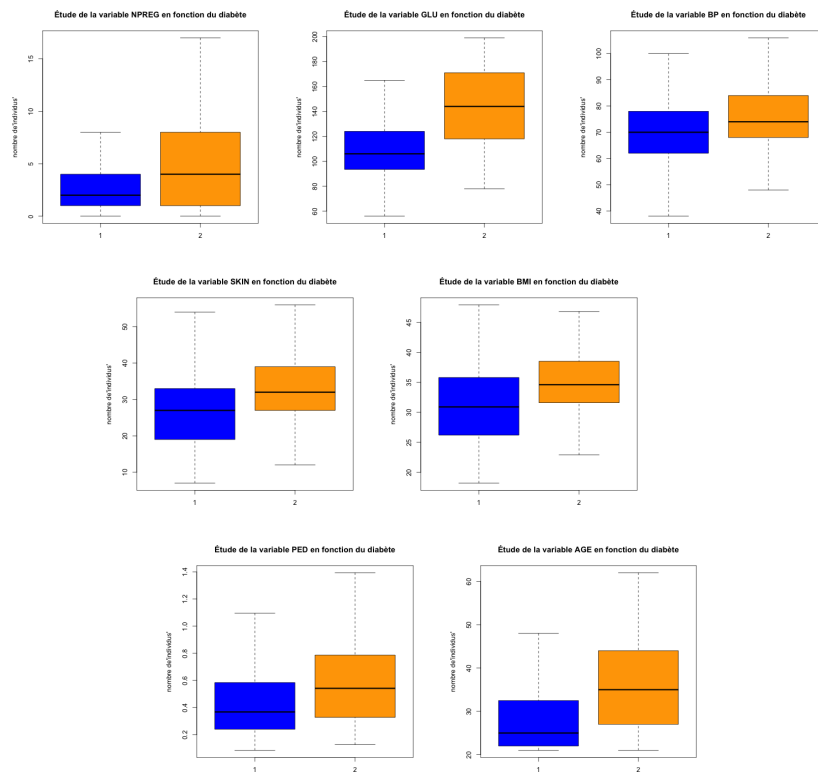


FIGURE 1.11 – Boxplot des données quantitatives

### 1.4.2 Analyse descriptive des données

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



On remarque une grande différence de distributions en fonction de la catégorie z. Les intervalles de confiance ne se chevauchent pas. La dispersion des données est aussi différente selon chaque variable. De ce fait, le diabète impacte

**Lien entre les variables**

## Chapitre 2

# Analyse Composantes Principales