Compte-rendu du TP01 de SY09
Statistique descriptive et Analyse en composantes principales

NGO Sy-Toan, Elliot BARTHOLME

6 avril 2017

1 Introduction

Dans cette section de TP, nous allons découvrir plusieurs jeux des données en les observant et en les analysant avec la méthode d'analyse en composantes principales (ACP).

Il y a 3 sets des données qui seront utilisés ; des notes de SY02 (UV de Statistiques) au printemps 2016, des données de Crabes (caractéristiques morphologiques des différentes espèces et sexes des crabes) et Pima, qui décrit les relations entre des femmes et le diabète via plusieurs caractéristiques. Après avoir décrit les variables et leurs propriétés (statistiques descriptives), nous passerons à l'étape suivante de la visualisation des ensembles de données avec l'analyse en composantes principales.

2 Statistique descriptive

2.1 Notes

Le jeu de données Notes donne des informations (domaine d'étude, niveau, diplôme...) sur les étudiants inscrits à l'UV SY02 au printemps 2016. Il fournit également les notes obtenues pendant le semestre et les correcteurs associés. Nous avons 296 étudiants, soit 296 observations avec 11 variables.

2.1.1 Analyse unidimensionnelle

- Specialité : Les branches sous-représentées sont le Génie des Procédés (GP), les Hutech (Humanités et Technologie) et les ISS (Master)
- Niveau : La plupart des étudiants est en niveau 2, soit GX02 pour les étudiants en branche du cursus ingénieur.
- Statut : A part 6 étudiants en échange, tout le monde est inscrit à l'UTC.
- Diplome : Les étudiants viennent à proportions à peu près égales pour la plupart de BAC ou DUT.
- Notes médian : La moyenne du médian est 10.92, la médiane 11 et l'étendue est maximale, 20.
- Correcteur médian : Le correcteur numéro 3 n'a corrigé aucun médian. Le reste est équitablement réparti sauf pour les correcteurs numéros 1 et 8 qui en ont corrigé deux fois moins que les autres.
- Notes final : La moyenne du final est 12.38, la médiane est 13 et l'étendue est de 19.50.
- Correcteur final : Le correcteur numéro 2 n'a corrigé aucun final, et les correcteurs numéros 1 et 8 en ont corrigé deux fois moins que les autres.
- Note totale : La moyenne totale (pondérée) sur les deux examens est de 11.84, la médiane est 12.30 et l'étendue est de 17.5.
- Résultat : 49 étudiants n'ont pas eu l'UV. Voici les fréquences de chaque résultat des étudiants :

	freq	percentage
F	49	17.25
Fx	34	11.97
E	37	13.03
D	44	15.49
C	58	20.42
B	42	14.79
A	20	7.04

FIGURE 1 – Fréquence de chaque résultat à l'UV dans l'échantillon d'étudiants

Nous pouvons résumer la description de chaque variable dans le tableau qui suit pour plus de clarté. La taille représente le nombre de valeurs possibles du domaine d'une variable (si dénombrable). Cela correspond donc à la cardinalité du domaine (pour les notes il y a 41 possibilités de 0,0.5,1,1.5... à 20, il y a 296 étudiants donc indices de Etu_i, 12 types de diplômes différents etc.).

	Nature	Domaine	Taille	Autres
Nom	Qualitative nominale	$\text{Etu}_i : i \in (1...296)$	296	
Spécialité	Qualitative nominale	{GP, GSU, GM, GI, GB, GSM, TC, HUTECH, ISS}	9	
Niveau	Quantitative discrète	{1...6}	6	
Statut	Qualitative nominale	{UTC, Echange}	2	
Diplome	Qualitative nominale	{BAC, DUT, autre 1er Cycle, CPGE, autre 2e cycle, autre diplôme supérieur, BTS, étranger supérieur, DEUG, Ingénieur, étranger secondaire, LICENCE}	12	Valeurs manquantes : 6 Explication : Etudiants en échange dont le diplôme n'est par conséquent pas caractérisé
Notes Médian	Quantitative discrète avec pas de 0.5	{0...20}	41	Valeurs manquantes : 3 Explication : Absence/Perte de copie /accident examen
Correcteur Médian	Qualitative nominale	$\text{Cor}_i : i \in \{1...8\}$	8	Valeurs manquantes : 3 Explication : pas de correcteur pour les étudiants sans médian
Notes Final	Quantitative discrète avec pas de 0.5	{0...20}	41	Valeurs manquantes : 12 Explication : Absence/Perte de copie /accident examen
Correcteur Final	Qualitative nominale	$\text{Cor}_i : i \in \{1...8\}$	8	Valeurs manquantes : 12 Explication : pas de correcteur pour les étudiants sans final
Note Totale	Quantitative discrète avec pas de 0.5	{0...20}	41	Valeurs manquantes : 12 Explication : étudiants non notés car médian ou final non noté
Résultat	Qualitative ordinale	{A, B, C, D, E,FX, F}	7	Valeurs manquantes : 12 Explication : relative à la note totale (pas de note donc pas de résultat)

FIGURE 2 – Tableau récapitulatif des variables étudiées

Nous avons décidé de retirer les 12 étudiants qui n'ont pas eu de note finale puisque ceux-ci ne pourrons être pris en compte lors des traitements numériques et ne participeront pas à la recherche de liens et corrélations entre les variables.

2.1.2 Analyse multidimensionnelle

2.1.3 Liens entre correcteurs et notes aux examens

Un correcteur étant un être humain, sa notation peut être légèrement subjective et c'est donc un facteur qui peut éventuellement influencer sur la note des étudiants. Nous allons examiner la relation entre les correcteurs et les notes aux examens, en distinguant les deux cas que sont l'examen médian et l'examen final.

Examen Median

Nous allons représenter graphiquement pour chacun des correcteurs la répartition des notes qu'il a données au médian. Le graphique des boîtes à moustaches associé est le suivant :

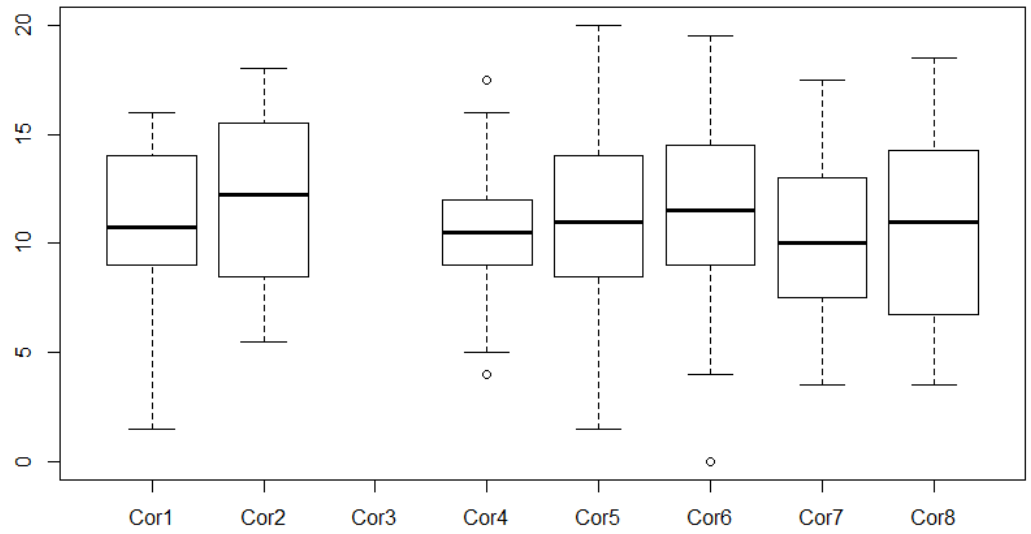

FIGURE 3 – Notes données au médian par chaque correcteur

La distribution est quelque peu variable mais ne semble pas influencer tant que ça les notes aux médians. Voici la moyenne de chaque correcteur numérotés de 1 à 8, où l'on voit que le correcteur 2 se démarque légèrement mais c'est tout. On a une variance de 0.39 autour de la moyenne de ce vecteur (chaque correcteur) et une étendue à 1.78 :

```
          [,1]     [,2] [,3]     [,4]     [,5] [,6]     [,7]     [,8]
[1,] 10.70833 12.01042  NaN 10.51111 11.26596 11.5 10.22917 10.76087
```

FIGURE 4 – Moyenne des notes données au médian par chaque correcteur de 1 à 8

On peut ensuite représenter sur une table de contingence les couples (correcteur.median,note.median) afin d'obtenir une distribution entre les deux variables. Nous pouvons alors faire un test du χ^2. Cependant, la condition n>5 doit au moins être vérifiée pour plus de 80% des cases de notre tableau, afin que l'échantillon soit suffisamment représentatif pour chaque critère. Ce n'est pas le cas puisque nous avons une matrice (7,37) où chaque note est décomposée. Nous décidons alors de réduire ce tableau en quatre catégorie de notes qui répartissent de manière uniforme les effectifs en fonction des correcteurs. On passe à une matrice (7,4) avec les étendues de notes suivantes :

1. *0 - 8*
2. *8.5 - 10.5*
3. 11 - 13.5
4. 14 - 20

Ce n'est pas une répartition basée sur des étendues de valeurs de taille égales, mais plutôt avec un volume uniforme de données (celles-ci étant regroupées autour de 10 les étendues sont plus strictes à ce niveau). Voici les moyennes de chacun de nos groupes de notes, où l'on a un écart maximal de 1, l'uniformité n'est donc pas mauvaise.

```
          [,1]     [,2]     [,3]     [,4]
[1,] 9.857143 9.714286 10.14286 10.85714
```

FIGURE 5 – Moyenne de chaque colonne (étendues de notes) de la table de contingence condensée

On applique maintenant le test sur la table (7,4). Nous formulons l'hypothèse nulle que le correcteur n'influe pas sur les notes obtenues au médian.

Test du χ^2 d'indépendance de Pearson pour le médian :

$$p - value = 0.4643 > 0.05 \tag{1}$$

Interprétation

Pour un taux à jusqu'à 45%, l'hypothèse nulle d'indépendance n'est pas rejetée. C'est-à-dire que même en prenant pratiquement un risque sur deux de se tromper en disant que les variables sont liées, on ne peut le garantir. On est largement au-dessus du niveau de confiance recommandé de 5%. Nous pouvons donc décider sans trop de risques que l'hypothèse nulle est acceptée : il n'existe pas de lien entre le correcteur du médian et la note obtenue.

Examen Final

Nous opérons de la même manière pour les notes du final. Nous passons d'une matrice M(7,36) à une matrice M(7,4) qui vérifie les conditions d'application du test du χ^2 d'indépendance (le logiciel R indique quand les conditions de ne sont pas respectées et que le résultat du test risque donc d'être biaisé).

Test du χ^2 d'indépendance de Pearson pour le final :

$$p - value = 0.2347 > 0.05 \tag{2}$$

Interprétation

On ne rejette pas de nouveau l'hypothèse d'indépendance entre les correcteurs et les notes du final, cependant à un taux de 25% on pourrait la rejeter. Il semble y avoir une plus forte corrélation entre les correcteurs et les notes du final que celles du médian. Cela est confirmé par le graphique des *boxplot* où l'on voit qu'il y a bien une plus forte variation entre les différents correcteurs :

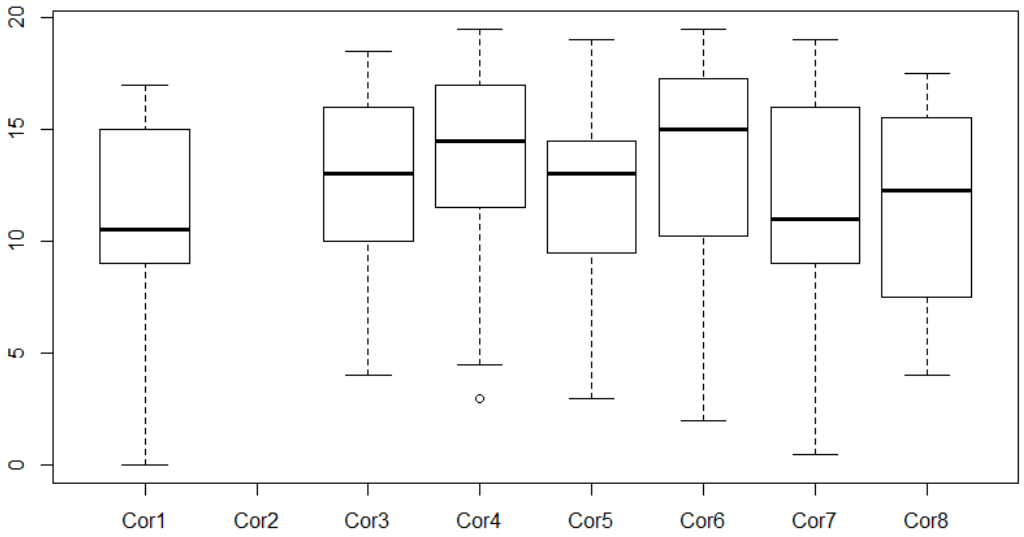

FIGURE 6 – Notes données au final par chaque correcteur

Ci-dessous le vecteur des moyennes des notes attribuées par chaque correcteur, avec une variance cette fois-ci de 0.93 et une étendue à 2.49 (variation donc plus grande que pour le médian) :

```
      [,1] [,2]     [,3]     [,4]     [,5]     [,6]     [,7]     [,8]
[1,] 10.94  NaN 12.57292 13.43478 11.82979 13.41489 11.90426 11.39583
```

FIGURE 7 – Moyenne des notes données au final par chaque correcteur de 1 à 8

2.1.4 Liens entre note median, note au final et note totale

Le lien entre ces variables est trivial puisqu'il découle de la relation linéaire entre la note attribuée à l'UV et les notes aux examens. La note totale est est une pondération de la note au médian et de celle au final selon des coefficients, qui ne changent pas. Le résultat à l'UV est beaucoup plus corrélé au final qu'au médian en raison de la pondération (coefficient) supérieure de l'examen. Les valeurs de corrélations suivantes le prouvent :

$$cor(note.median, note.totale) = 0.75 \tag{3}$$

$$cor(notes.final, note.totale) = 0.92 \tag{4}$$

$$cor(notes.final, note.totale) > cor(notes.median, note.totale) \tag{5}$$

Regardons pour finir la corrélation entre la note du médian et la note du final. Autrement dit, est-ce qu'un élève aura le même profil au médian et au final (s'il rate le médian par exemple, ratera-t-il le final aussi) ?

$$cor(note.median, note.final) = 0.43 \tag{6}$$

Celle-ci est loin d'être significative, et cela se vérifie à maintes reprises dans les données : de nombreux élèves ayant raté le médian rebondissent au final, mais l'inverse se produit aussi, certains étudiants ayant obtenu de très bonne notes au médian échouent le final et obtiennent même une note éliminatoire pour certains.

2.1.5 Liens entre spécialité et note finale

Il est maintenant intéressant de regarder l'influence d'une spécialité sur la note totale de l'UV. Si l'on produit les *boxplot* des notes totales en fonction de la spécialité on voit une disparité plutôt significative. Ce résultat doit cependant être pondéré par le fait que les effectifs sont loin d'être égaux (un seul étudiant en Hutech par exemple...). Les TC semblent mieux réussir que les branches de manière générale. On pourrait aussi tracer les *boxplot* en fonction du résultat (lettres F et Fx correspondent à une non-obtention de l'UV) et l'on obtiendrait le même schéma.

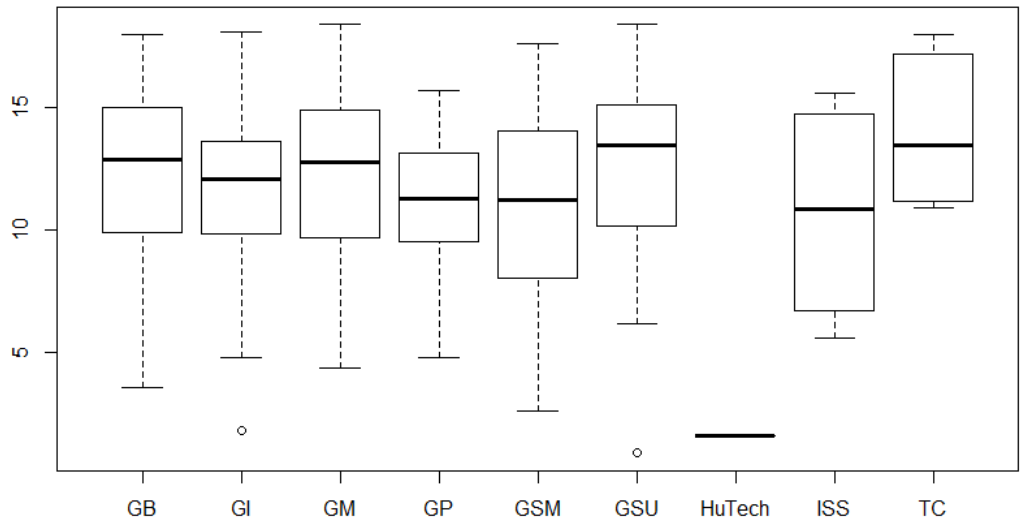

FIGURE 8 – Notes totales en fonction de la spécialité de l'étudiant

Il est difficile de mettre en place les test d'indépendances du χ^2 car nos effectifs ne sont pas assez grands, même en regroupant les notes en quatre groupes. On peut utiliser le test exact de Fisher sur une petite matrice comme par exemple celle contenant les branches Hutech, ISS et TC seulement, où l'on obtient une $p - value = 0.3566$; celle-ci est cependant biaisée puisque l'on a considéré seulement trois branches pour les liens d'indépendance. Un test du χ^2 sur une matrice réduite à 4 étendues de notes (et toujours les 9 branches) donne une $p-value = 0.1971$.

Dans tous les cas on semble accepter l'hypothèse d'indépendance mais avec des tests dont les conditions d'application ne sont pas totalement remplies. Il est donc dur de conclure et nous dirons que le seul lien existant est avec les Tronc Commun (TC) ceux-ci réussissent bien l'UV de manière générale, peut-être car c'est une UV de branche qui est donc prise soit par des étudiants en avance (bons élèves) soit par des étudiants en retard qui font un tronc commun plus long mais qui ont donc plus d'expérience que des TC *classiques*.

2.1.6 Liens entre niveau et note finale

Il peut également être intéressant d'observer la réussite des élèves en fonction de leur niveau. On doit cependant considérer les étudiants de TC et de branches indépendamment car, par exemple, un niveau 4 TC ne correspond pas à un niveau 4 de branche. De plus, on se rend compte que seuls les TC de niveau 4 ou 5 ont choisi SY02 (ce qui est dû au prérequis de cette UV). Le niveau n'apportera donc pas d'information pertinente pour leur cas et nous avons choisi de les délaisser pour cette partie. L'influence de cette spécialité est étudiée dans la partie précédente. Nous nous concentrons sur le niveau en branche.

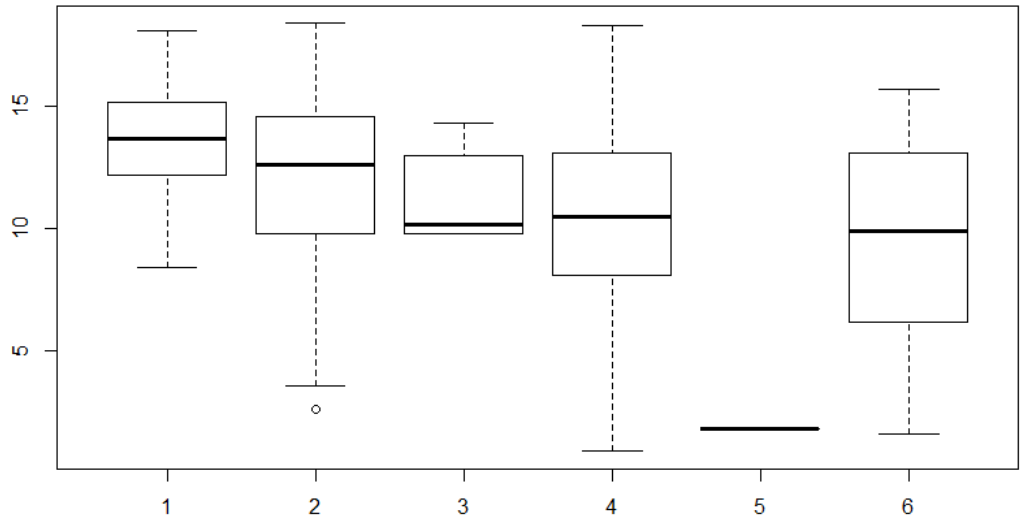

FIGURE 9 – Notes totales en fonction du niveau des étudiants en branche

Le niveau semble bien avoir une influence sur les notes obtenues à l'UV. On voit en effet une décroissance des *boxplot* au fur et à mesure que les niveaux augmentent. Cela signifie que plus l'étudiant est ancien, plus ses résultats à l'UV décroissent. C'est à nouveau pondéré car le niveau 2 est sur-représenté (154 d'effectif) et les niveaux 3,5,6 sous-représentés (effectifs de 5,1 et 9).

Nous effectuons maintenant un test d'indépendance du χ^2. On produit la table de contingence entre le niveau de l'étudiant en branche et la note totale obtenue. On regroupe les colonnes (notes) en 4 étendues comprenant chacune un effectif similaire. Les étendues sont les suivantes :

1. *0.9 - 9.6*
2. *9.7 - 12.2*
3. *12.3 - 14.7*
4. *14.8 - 18.4*

Voici l'effectif pour chaque catégorie de notes :

```
     [,1] [,2] [,3] [,4]
[1,]   68   68   71   68
```

FIGURE 10 – Effectif de chacune des catégories des notes qui ont été fusionnées

On a donc une répartition qui est bien uniforme. Nous devons cependant maintenant regrouper aussi les lignes car les effectifs ne sont toujours pas satisfaisants pour chaque contingence afin de remplir les conditions du test. Si l'on effectue le test avec cette table, on obtient une $p - value = 0.003028$. Nous choisissons alors de regrouper les niveaux en trois catégories au lieu de 6 :

— Début de branche : niveaux 1-2
— Milieu de branche : niveaux 3-4
— Fin de branche : niveau 5-6

Nous sommes donc finalement passés d'une matrice de contingence (6,115) à une matrice (3,4).

Test du χ^2 d'indépendance :

$$p - value = 0.01146 < 0.05 \tag{7}$$

Interprétation

L'hypothèse nulle est donc rejetée pour un niveau de 95% de confiance. Il semble alors bien exister un lien statistique entre le niveau de l'étudiant de branche et sa note totale à l'UV.

2.1.7 Lien entre origine de l'étudiant, dernier diplôme obtenu et note à l'UV

Nous allons maintenant rapidement étudier la réussite de l'UV en fonction de l'origine de l'étudiant.

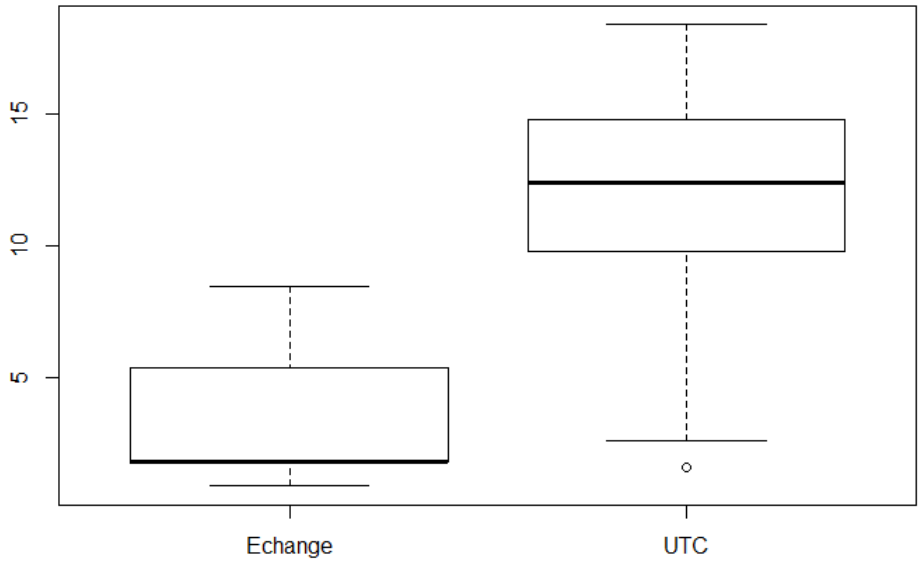

FIGURE 11 – Notes totales en fonction de l'origine de l'étudiant

Selon la lettre obtenue :

```
> notes[notes$statut=="Echange",]$resultat
[1] F  Fx F  F  F
```

FIGURE 12 – Lettres obtenues par les étudiants étrangers

On voit que les étudiants en échange ont de mauvais résultats. Aucun n'a réussi l'UV (un a été absent, il n'apparaît pas sur les schémas). Regardons maintenant en fonction du dernier diplôme obtenu :

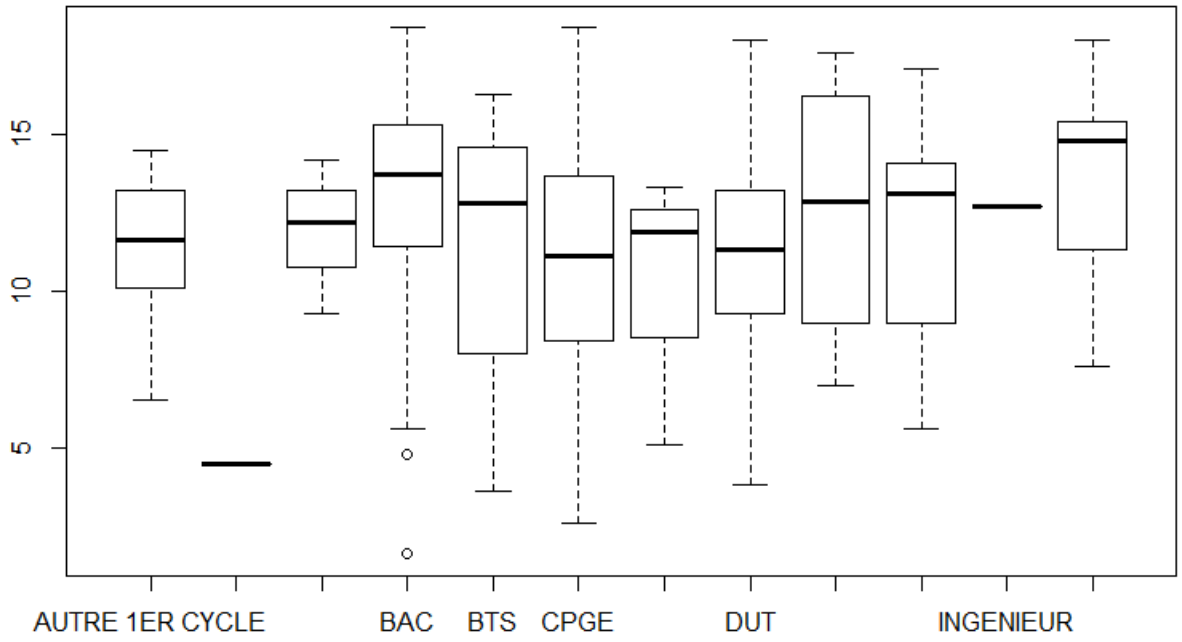

FIGURE 13 – Notes totales obtenues selon le dernier diplôme obtenu

Le dernier diplôme obtenu a clairement une influence sur les notes à l'UV et donc la réussite. Les étudiants venant de BAC et de licence (dernier *boxplot*) réussissent en moyenne mieux que les autres. Les dispersions de données dans chaque boîte sont très éparses selon le diplôme d'origine, c'est donc un facteur important.

2.1.8 Remarque

Dans la plupart des observations nous avons utilisé la note totale et non le résultat à l'UV car la note totale est numérique et donc plus significative pour le lecteur et lors de calculs ou de tracés de graphiques.
De plus, le résultat peut parfois ne pas être en relation avec la note totale en fonction du jury de l'UV qui peut décider différemment selon le comportement de l'étudiant durant le semestre (relâchement, absentéisme...) ; il y a par exemple un étudiant qui a eu A avec 16.4 de moyenne et un autre qui a eu B avec 16.5 de moyenne.
Finalement les notes ≤ 6 au final sont éliminatoires et entraîneront immédiatement une non-obtention (exemple d'un élève ayant eu 16 au médian et 6 au final soit 10 de moyenne mais un F car 6 est éliminatoire au final).

2.2 Données crabs

Description des données

Nous traitons ici un jeu de données sur des crabes de deux espèces différentes : orange et bleue. On a 200 observations, 100 pour chacune des espèces. Chacune des espèces est divisée en deux parties selon le sexe, soit quatre groupes au total : Mâle (M) ou Femelle (F) et Orange (O) ou Bleu(B). On a finalement 50 mâles bleus, 50 mâles oranges, 50 femelles bleues, 50 femelles oranges. Nous allons tenter de répondre à deux problématiques :

— Est-il possible de d'identifier l'espèce ou le sexe d'un crabe à partir de plusieurs caractéristiques morphologiques ?
— Quelle est la corrélation entre les variables, et pourquoi ?

Observons de plus près les variables de l'échantillon de 200 crabes :

Nom	Nature	Domaine	Taille
SP	Qualitative nominale	{O, B}	2
Sex	Qualitative nominale	{F, M}	2
Index	Qualitative ordinale	{1 ... 5}	50
FL	Quantitative continue	$\mathbb{R}$	infini
RW	Quantitative continue	$\mathbb{R}$	infini
CL	Quantitative continue	$\mathbb{R}$	infini
CW	Quantitative continue	$\mathbb{R}$	infini
BD	Quantitative continue	$\mathbb{R}$	infini

FIGURE 14 – Description des données Crabs

Regardons maintenant les résumés numériques selon chacun des deux critères : le sexe et l'espèce.

```
     FL      RW      CL      CW      BD
15.432 13.487 31.360 35.830 13.724
> colMeans(crabsmalequant)
     FL      RW      CL      CW      BD
15.734 11.990 32.851 36.999 14.337
```

FIGURE 15 – Moyenne de chacune des variables morphologiques des crabes femelles puis mâles

On voit que les crabes mâles sont en moyenne plus gros sauf au niveau de l'arrière où ils sont moins larges que les femelles. Cependant, les différences sont loin d'être significatives et ne permettent pas d'identifier clairement un individu par rapport à un autre.

```
> colMeans(crabsbluequant)
     FL      RW      CL      CW      BD
14.056 11.928 30.058 34.717 12.583
> colMeans(crabsorangequant)
     FL      RW      CL      CW      BD
17.110 13.549 34.153 38.112 15.478
```

FIGURE 16 – Moyenne de chacune des variables morphologiques des crabes bleus puis oranges

On voit que les crabes oranges sont en moyenne plus gros que les crabes bleus. La différence est ici plus significative. Traçons les *boxplot* pour chaque caractéristique :

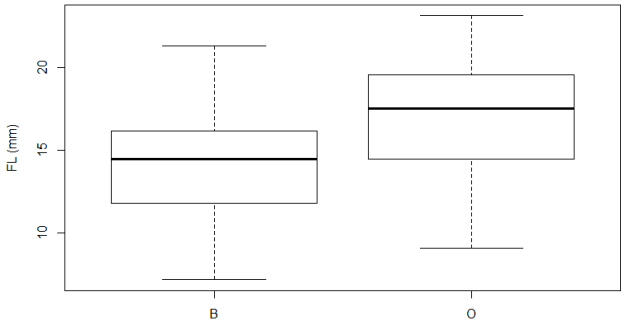

FIGURE 17 – Taille du lobe frontal en fonction de l'espèce

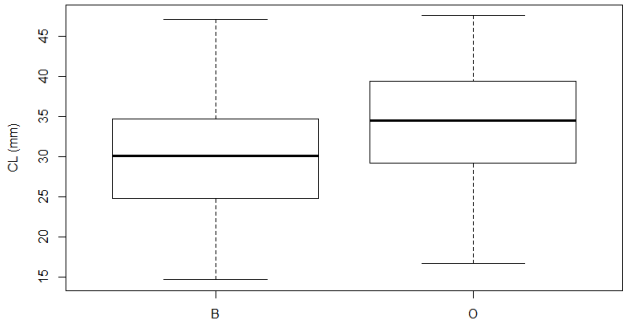

FIGURE 18 – Longueur de la carapace en fonction de l'espèce

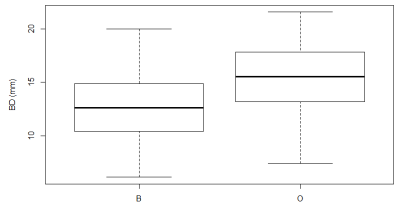

FIGURE 19 – Profondeur du corps en fonction de l'espèce

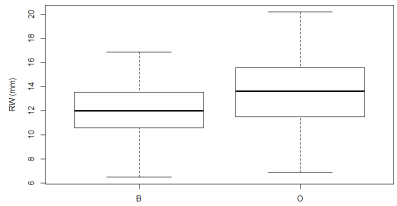

FIGURE 20 – Largeur de l'arrière en fonction de l'espèce

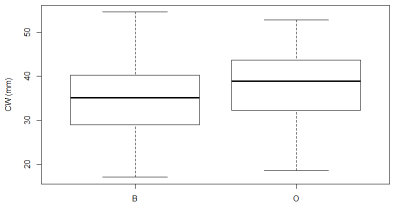

FIGURE 21 – Largeur de la carapace en fonction de l'espèce

Ces différents graphiques confirment cette hypothèse : les crabes orange sont plus gros en moyenne que les bleus.

Traçons maintenant le graphique matriciel des caractéristiques de crabes selon leur espèce ou leur sexe afin d'étudier les liens entre variables :

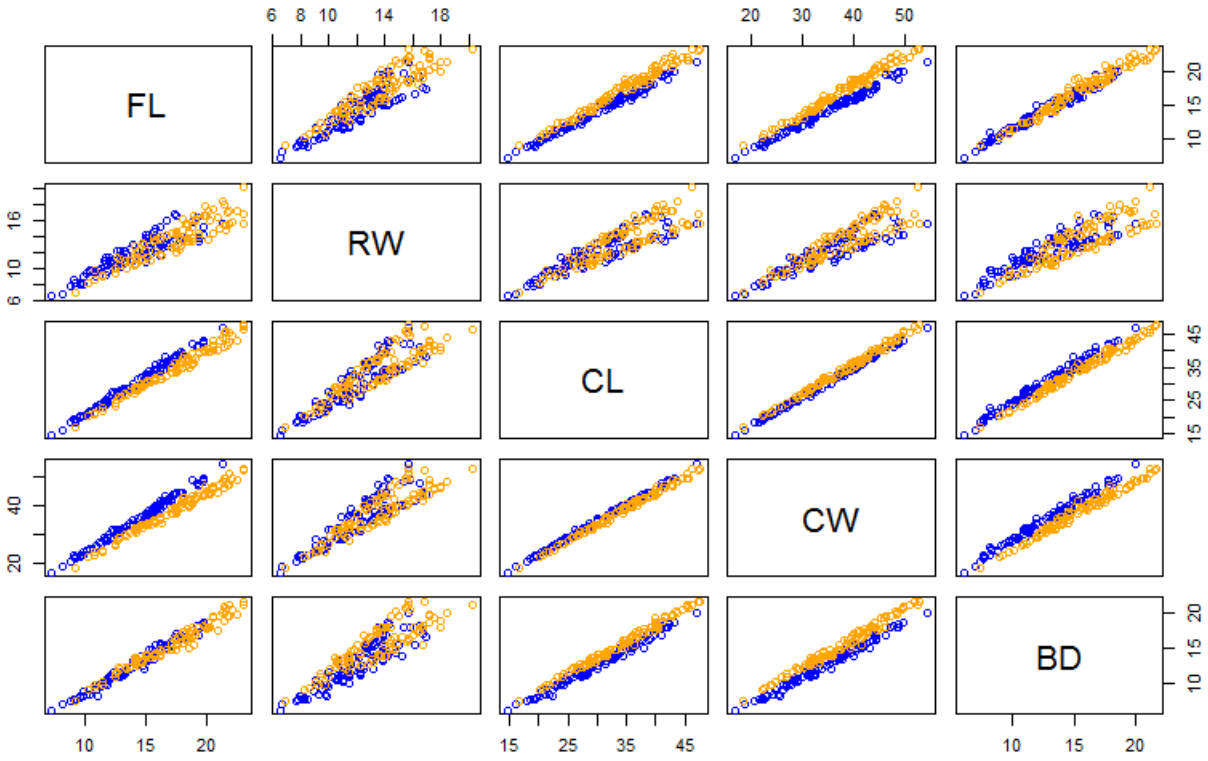

FIGURE 22 – Représentation des caractéristiques des membres de crabes selon leur espèce (couleurs représentatives)

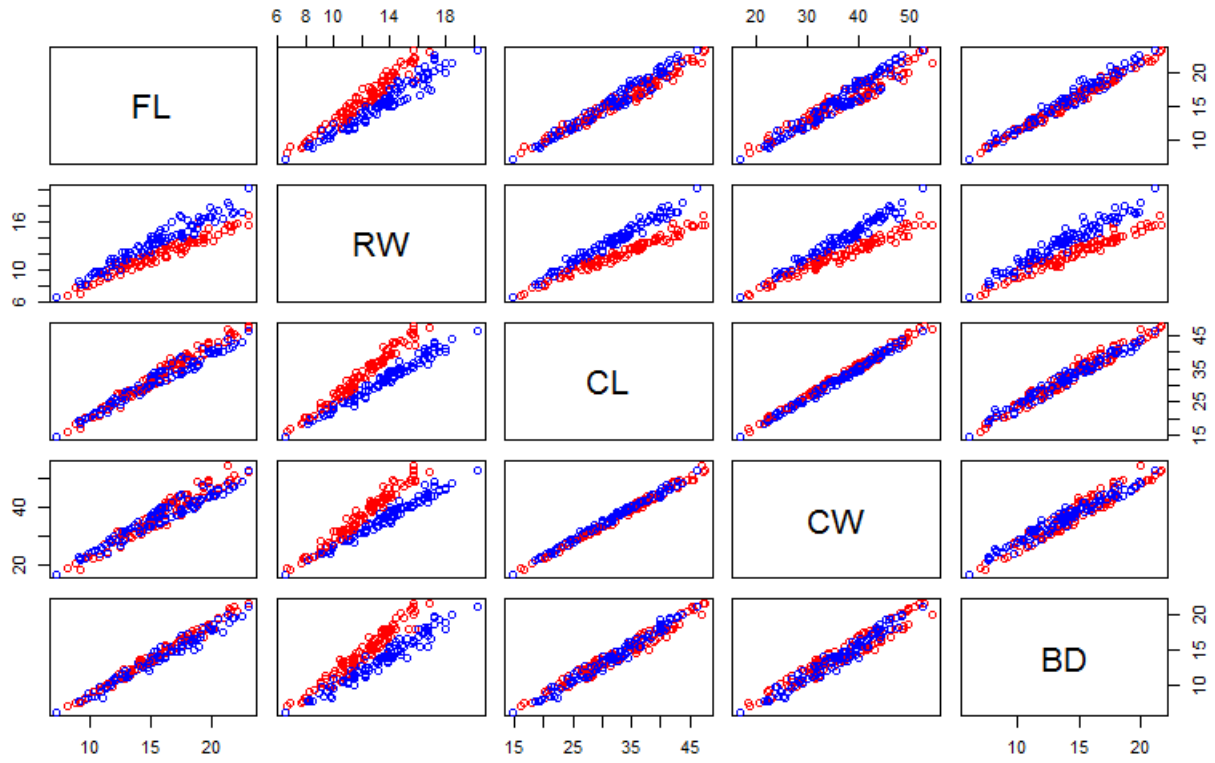

FIGURE 23 – Représentation des caractéristiques des membres de crabes selon leur sexe *(rouge = mâle, bleu = femelle)*

Lors de la distinction des individus par sexe, on observe que, contrairement à la distinction par espèce, certaines paires de caractéristiques morphologiques et donc les rapports entre les deux, font apparaître des différences selon le mâle ou la femelle. Le ratio RW/CL et RW/CW est par exemple significativement plus grand chez les femelles que chez les mâles.
Regardons en détail le graphique entre RW et CL des crabes mâles et femelles :

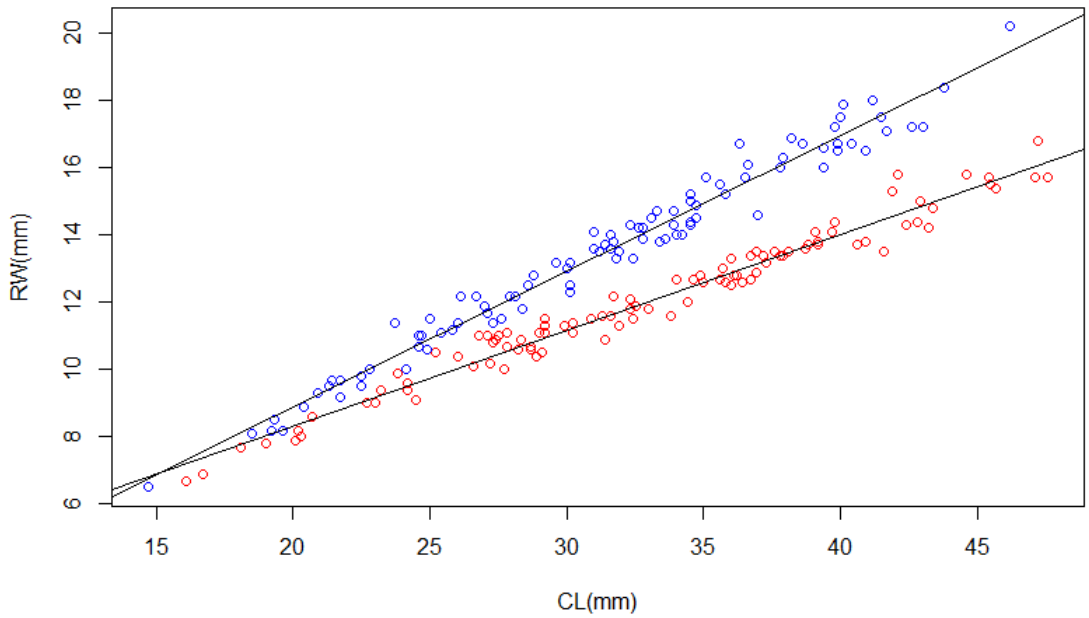

FIGURE 24 – Relation entre la largeur arrière et la longueur de la carapace entre mâles et femelles

Le coefficient directeur de la droite de régression des femelles est 0.40 et celui des mâles 0.28.
Ainsi, pour certains caractéristiques, il semble possible en étudiant leur rapport (ratio) de déterminer le sexe de l'individu.

Étude des corrélations
Voici la matrice des corrélations entre les caractéristiques des crabes :

	FL	RW	CL	CW	BD
FL	1.0000000	0.9069876	0.9788418	0.9649558	0.9876272
RW	0.9069876	1.0000000	0.8927430	0.9004021	0.8892054
CL	0.9788418	0.8927430	1.0000000	0.9950225	0.9832038
CW	0.9649558	0.9004021	0.9950225	1.0000000	0.9678117
BD	0.9876272	0.8892054	0.9832038	0.9678117	1.0000000

FIGURE 25 – Matrice des corrélations des données numériques de Crabs

Comme on peut le voir, quasi toutes les corrélations sont à 1. C'est-à-dire que toutes les caractéristiques des crabes sont très corrélées, et ce de manière linéaire. Bien sûr cela se voyait déjà sur les graphiques matriciels 22 et 23. Quelle en est la cause ?

Celle qui semble la plus évidente est que les caractéristiques physiques des et individus d'une espèce animale sont très presque toujours corrélées positivement : si l'individu est gros, tous ses membres seront en moyenne plus gros que l'individu plus petit. Il apparaît logique que le crabe qui a un lobe frontal très gros n'ait pas une carapace toute petite.

Comment remédier à cela ?
Une solution peut être, au lieu de comparer les tailles en valeur absolue, de comparer les proportions de chaque caractéristique/membre par rapport à la taille totale de l'individu. On opère alors une *pondération* à chaque valeur en la divisant par somme de la taille de tous les membres du crabe (somme des lignes).

```
plot(crabsquant/rowSums(crabsquant), col=c("blue","red")[crabs$sex])
```

De plus, cela permet d'y voir plus clair lors du cas suivant :
Si l'on prend deux crabes, l'un orange et l'autre bleu par exemple, si le orange est tout jeune et le bleu très âgé, ce dernier sera sûrement plus gros que le orange malgré ce que l'on a indiqué précédemment. La proportion de chaque membre par rapport à sa taille permettra alors de d'en déduire plus facilement l'espèce.

Voici les deux graphes matriciels des caractéristiques une fois pondérées par la taille totale de l'individu.

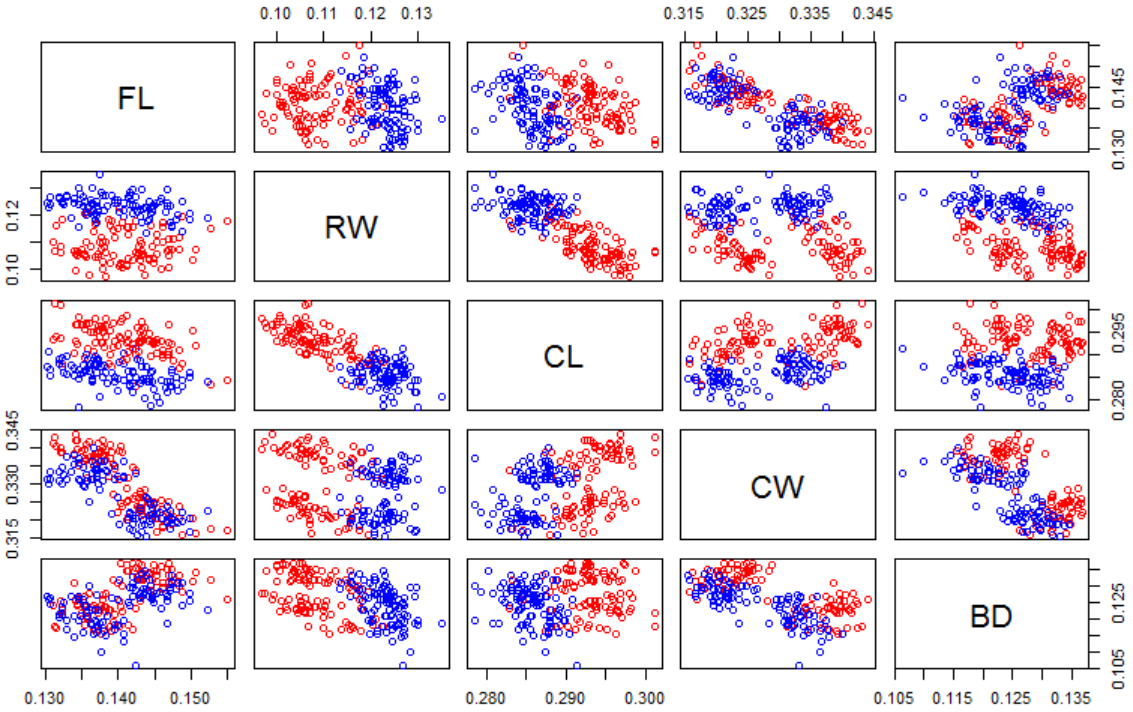

FIGURE 26 – Graphique matriciel des caractéristiques des crabes selon le sexe, mais en version pondérée

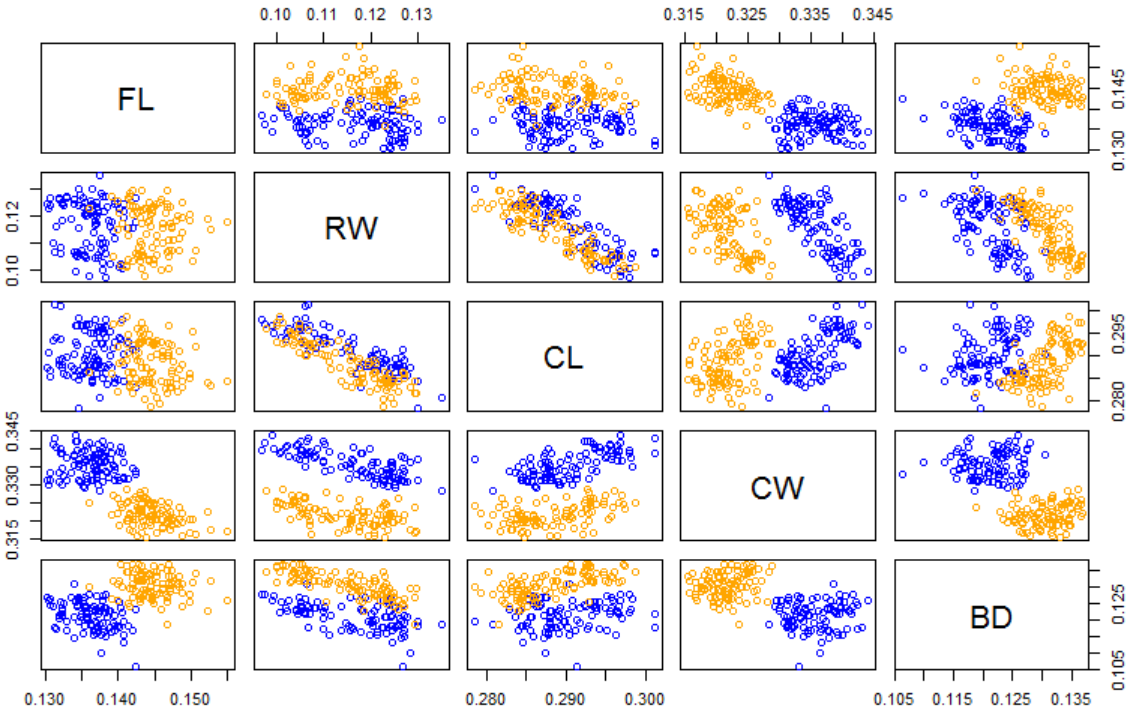

FIGURE 27 – Graphique matriciel des caractéristiques des crabes selon l'espèce, mais en version pondérée

Sur ces deux figures la distribution est bien différente qu'auparavant et il n'y a plus aucune corrélation. Pour les espèces on a clairement pour chaque rapport deux nuages bien distincts représentant chacune. Les proportions ne sont pas corrélées entre elles, ce qui est logique, mais on peut voir rapidement quelles caractéristiques séparent l'échantillon en deux. Par exemple, le lobe frontal est bien plus gros chez les crabes orange ($y_{orange} > y_{bleu}$ sur toutes les lignes ou FL est en ordonnée). On pourra étudier les proportions des crabes indépendamment du facteur âge qui influe sur la taille.

2.3 Données Pima

Nous traitons ici un jeu de données sur des individus de sexe féminin, diabétiques et non-diabétiques. On a 532 observations, 355 cas non-diabétiques et 177 cas diabétiques.
Nous allons dans un premier temps réaliser une analyse descriptive des données puis tenter d'identifier les liens statistiques forts entre variables ; notamment du facteur "diabète" et son influence sur les indicateurs numériques présents dans le jeu de données.

Nom	Nature	Domaine	Volume
npreg	Quantitative discrète	{0 ... 17}	infini
glu	Quantitative discrète	{56 ... 199}	infini
bp	Quantitative discrète	{24 ... 110}	infini
skin	Quantitative discrète	{7 ... 99}	infini
bmi	Quantitative continue	{18.20 ... 67.10}	infini
ped	Quantitative continue	{0.085 ... 2.420}	infini
age	Quantitative discrète	{21 ... 81}	infini
z	Qualitative nominale	{1,2}	2

FIGURE 28 – Description des données Pima

Analyse des liens statistiques
Regardons maintenant les résumés numériques selon le critère de diabète.

```
> colMeans(Pima[Pima$z==1,-8])
     npreg        glu         bp       skin        bmi        ped        age
 2.9267606 110.0169014  69.9126761  27.2901408  31.4295775  0.4463155  29.2225352
> colMeans(Pima[Pima$z==2,-8])
     npreg        glu         bp       skin        bmi        ped        age
 4.7005650 143.1186441  74.7005650  32.9774011  35.8197740  0.6165876  36.4124294
```

FIGURE 29 – Moyenne de chacune des variables des individus non-diabétiques et diabétiques

On voit que les individus diabétiques ont tous les paramètres en moyenne plus gros que les individus non-diabétiques, la différence de taux plasmatique de glucose (glu) est notamment la plus significative.

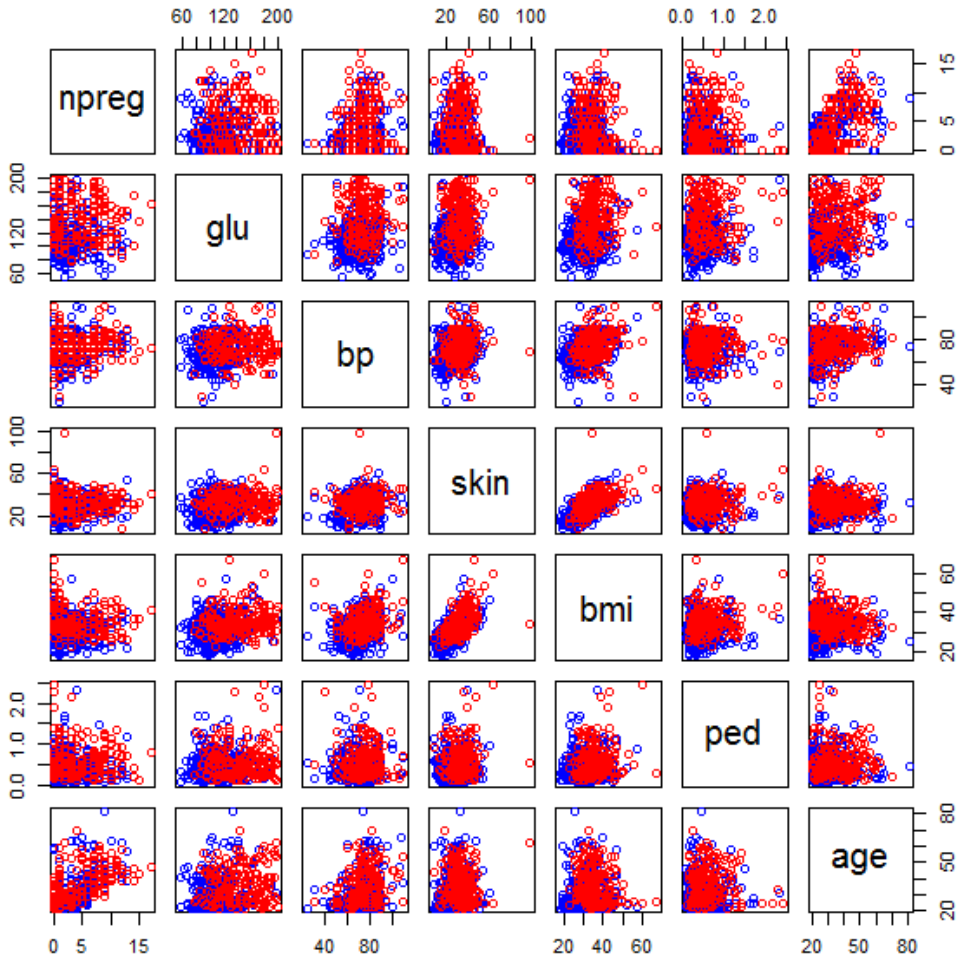

FIGURE 30 – Représentation des caractéristiques des individus selon diabète *(rouge = diabète, bleu = pas de diabète)*

Voici maintenant la matrice des corrélations en version graphique afin de voir rapidement les liens forts :

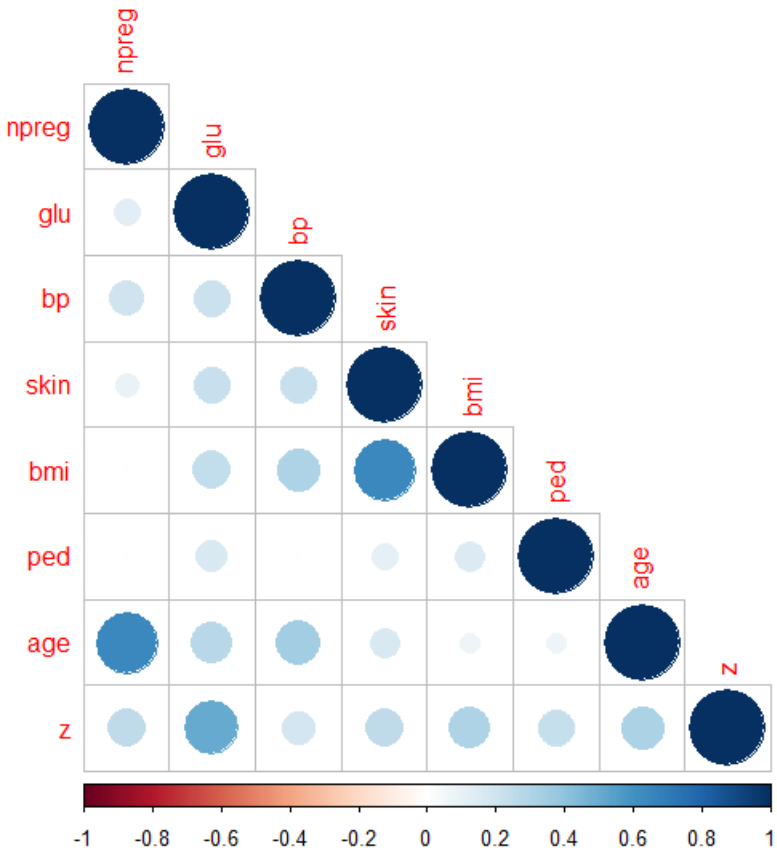

FIGURE 31 – Corrélation entre variables de Pima

Nous observons 3 liens statistiques forts :

— entre npreg (nombre de grossesses) et age : lien trivial puisque le nombre de grosses augmente forcément avec l'âge ;

— entre skin (épaisseur du pli cutané au niveau du triceps) et bmi (l'indice de masse corporelle) : le lien apparaît logique puisque les IMC élevé ont tendances à induire une masse grasse plus grande, sauf chez les grands sportifs ;

— entre z (indicateur du diabétisme ou non) et glu (taux plasmatique de glucose) : le diabète est une maladie qui se caractérise par une hyperglycémie chronique. Le malade n'arrive pas à auto-réguler son taux de glucose dans le sang, celui-ci est par conséquent en permanence trop élevé. Il est donc logique qu'un individu diabétique soit associé à un taux plasmatique de glucose plus élevé.

Il peut être intéressant de remarquer que l'absence de forte corrélation entre le *bmi* et z ainsi qu'entre l'age et z dénote très probablement une majorité de diabètes de type 1 (maladie liée à une absence de sécrétion d'insuline par le pancréas) dans notre échantillon, contrairement au type 2 qui est une maladie se développant au fil des années avec l'âge et à cause d'un fort indice de masse corporelle et de mauvaises habitudes alimentaires.

Bien évidemment, les corrélations ont été étudiées ici en considérant la variable z comme quantitative. Cela fonctionne car le diabète est caractérisé par la valeur la plus haute (2, relation d'ordre croissant) de z. Il serait cependant bien plus adéquat de la considérer en tant que qualitative comme nous l'avons définie au début.
On effectue un test de χ^2 d'indépendance entre z et *glu* afin de confirmer leur lien fort. Un tel test donne $p - value = 4.38e - 08 << 0.05 << 0.01$ ce qui confirme bien la forte dépendance entre ces deux critères pour des niveaux allant jusqu'à plus de 99% de confiance.

3 Analyse en composantes principales

3.1 Exercice théorique

3.1.1 Calculer les axes factoriels de l'ACP

Pour calculer les axes factoriels de l'ACP du nuage de points défini par les quatre variables quantitatives, on centre tout d'abord la matrice des correcteurs qui s'appelle X, et après on trouve la matrice de variance avec l'équation :

$$\frac{1}{n}X^T X$$

Les valeurs propres de la matrice de variance sont les valeurs d'inertie expliquée, et les vecteurs propres sont les axes factoriels de l'ACP.

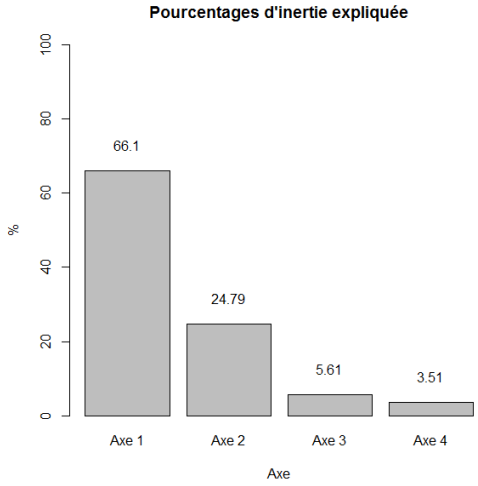

FIGURE 32 – Représentation les pourcentages d'inertie expliquée

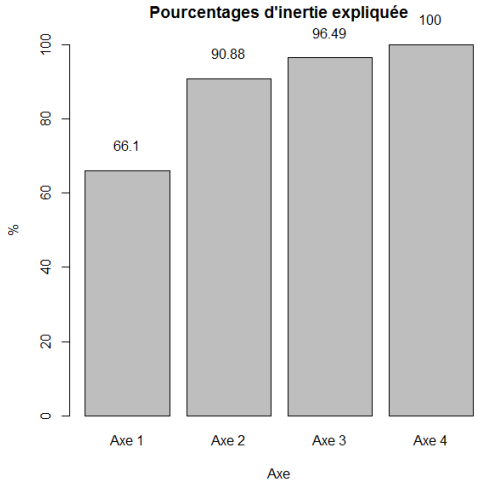

FIGURE 33 – Représentation des pourcentages d'inertie expliquée

En conclusion, l'axe 1 représente 66.1% d'inertie expliquée quand l'axe 1 et l'axe 2 représentent 90.88% d'inertie expliquée. Le nuage de points peut alors être représenté par ces deux axes sans perdre trop d'informations.

3.1.2 Calculer les composantes principales

Les composantes principales sont calculées grâce à l'équation :

$$C = XMU = XU$$

Avec $M = I_p$ (parce que on a le mêmes pondérations à tous les individus) et U est la matrice des vecteurs propres.

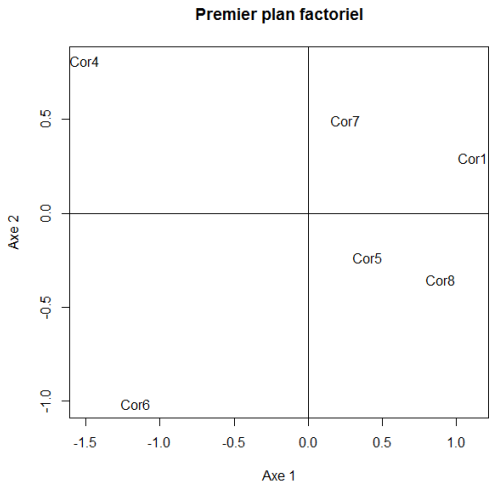

FIGURE 34 – La représentation des six individus dans le premier plan factoriel.

3.1.3 La représentation des quatre variables dans le premier plan factoriel

A présent, grâce au calcul des corrélations, nous pouvons les quatre variables dans le premier plan factoriel :

$$diag(\frac{1}{\sqrt{\frac{N}{N-1}} * apply(X, 2, sd)}) * U * \sqrt{L} \tag{8}$$

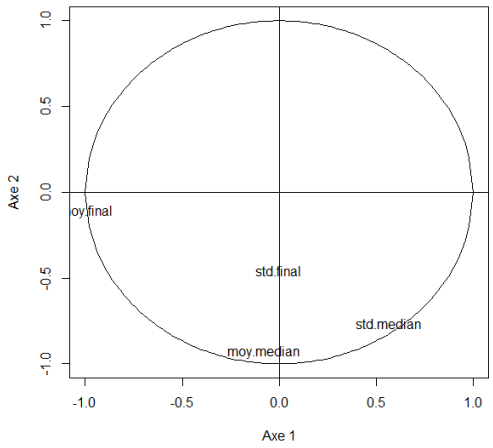

FIGURE 35 – La représentation des quatre variables dans le premier plan factoriel.

Nous pouvons donc conclure que l'axe 1 est principalement corrélé avec

Recomposition de la matrice

Calcul de l'expression $\sum_{\alpha=1}^{k} c_\alpha u_\alpha^T$: pour k=4 on obtient la matrice suivante X initiale des correcteurs, centrée en colonne. C'est une opération de *reconstitution* qui vise à retrouver la matrice X dont est issue l'ACP via les composantes principales et les axes factoriels associés.

3.1.4 La représentation des huit individus dans le premier plan factoriel.

Les données sont *nettoyées* en remplaçant les valeurs manquantes des deux correcteurs par la moyenne de la variable associée. On obtient la matrice suivante :

	moy.median	std.median	moy.final	std.final
[1,]	10.70833	3.900715	10.94000	4.583303
[2,]	12.01042	3.712385	12.15326	4.515389
[3,]	10.71418	4.056270	12.57292	3.648068
[4,]	10.23469	3.043268	13.43478	4.343077
[5,]	10.97959	4.413473	11.82979	3.971743
[6,]	11.50000	4.303584	13.41489	4.877097
[7,]	10.12245	4.030522	11.90426	4.444878
[8,]	10.74000	4.646056	11.39583	4.872235

FIGURE 36 – Jeu de données initial corrigé avec la moyenne

On ré-applique alors la méthode ACP détaillé précédemment avec ces données et on obtient une nouvelles représentation des correcteurs des huit correcteurs dans les deux premiers plans factoriels (trois axes).

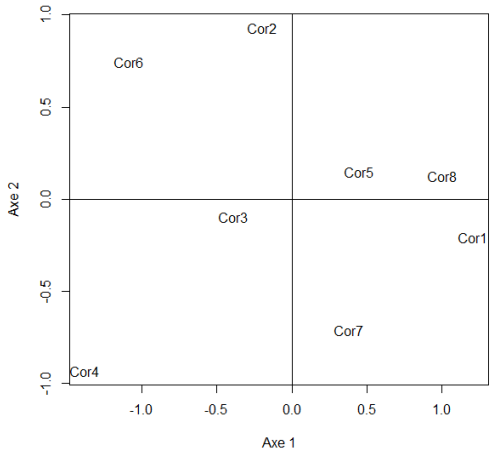

FIGURE 37 – Le plan de représentation 1-2

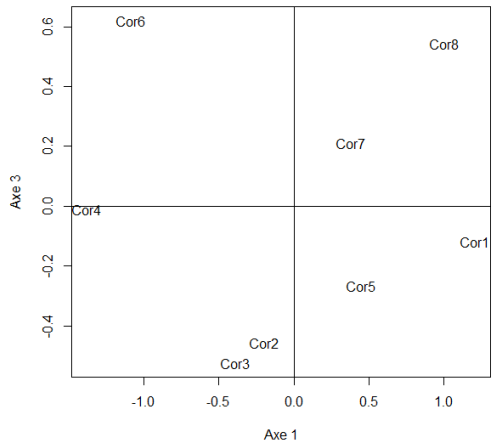

FIGURE 38 – Le plan de représentation 1-3

On retrouve la même répartition des correcteurs pour le premier plan factoriel *1-2*. C'est normal car l'ajout des deux correcteurs ne bouleversent pas la répartition puisqu'on leur a associé la moyenne, ils se retrouvent ainsi centrés dans le nuage de points des individus.

3.2 Utilisation des outils R

Regardons maintenant les outils proposés par R pour un effectuer une ACP sur des données :

```
A <- princomp(correcteurs) #effectuer l'ACP du jeu de données notes
U <- A$sdev^2    #les valeurs propres
L <- A$loadings    #les axes principaux
C <- A$scores    #les composantes principales
```

Les résultats obtenus sont les suivants :

```
> princomp(correcteurs[,-1])$sdev^2
   Comp.1    Comp.2    Comp.3    Comp.4
0.7605401 0.3618012 0.1620380 0.1197456
> princomp(correcteurs[,-1])$loadings

Loadings:
           Comp.1 Comp.2 Comp.3 Comp.4
moy.median          0.929 -0.291  0.205
std.median  0.291  0.343  0.453 -0.770
moy.final  -0.950         0.222 -0.221
std.final           0.137  0.813  0.562

               Comp.1 Comp.2 Comp.3 Comp.4
SS loadings      1.00   1.00   1.00   1.00
Proportion Var   0.25   0.25   0.25   0.25
Cumulative Var   0.25   0.50   0.75   1.00
> princomp(correcteurs[,-1])$scores
         Comp.1      Comp.2      Comp.3      Comp.4
[1,]  1.2088931 -0.20886399 -0.11944756  0.43169305
[2,] -0.1963305  0.93112310 -0.45706552  0.45485784
[3,] -0.3875706 -0.09613188 -0.52447852 -0.49752323
[4,] -1.3771814 -0.93446141 -0.01053229  0.30895922
[5,]  0.4483166  0.14988402 -0.26592460 -0.44784979
[6,] -1.0845403  0.74648814  0.62096484 -0.09718104
[7,]  0.3807964 -0.71159015  0.21167097 -0.07913283
[8,]  1.0076167  0.12355216  0.54481267 -0.07382322
```

FIGURE 39 – Résultats obtenus grâce à l'outil princomp de R

Les accesseurs permettent d'avoir les résultats obtenus dans la partie précédente.

Les fonctions biplot et plot :

— La fonction plot permet de visualiser sous forme de graphique les valeurs propres de chaque vecteur propre
— La fonction biplot permet de visualiser dans un même graphique la répartition dans le premier plan factoriel des individus et des variables.

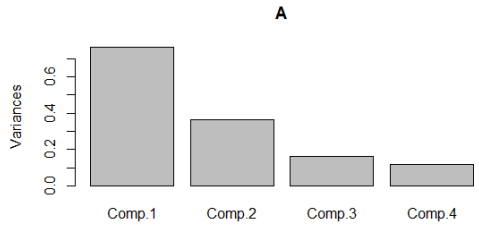

FIGURE 40 – Resultat de plot avec princomp objet

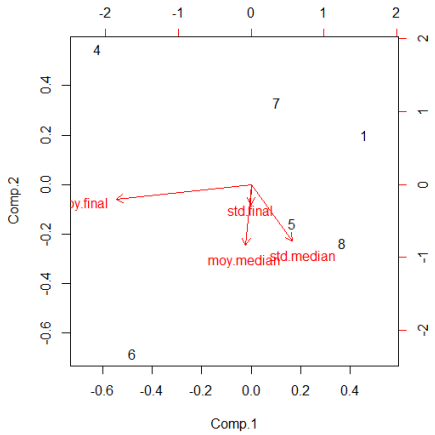

FIGURE 41 – Résultat de biplot avec princomp objet

On retrouve visuellement les mêmes résultats que dans l'ACP réalisé dans l'exercice 2.1. On remarque tout de même une différence dans les échelles. Cette différence peut s'expliquer par le fait que notre ACP était en réalité biaisée dans l'exercice 2.1.

3.3 Données Crabs

On applique une première ACP sur les données Crabs. On visualise les résultats avec *biplot* :

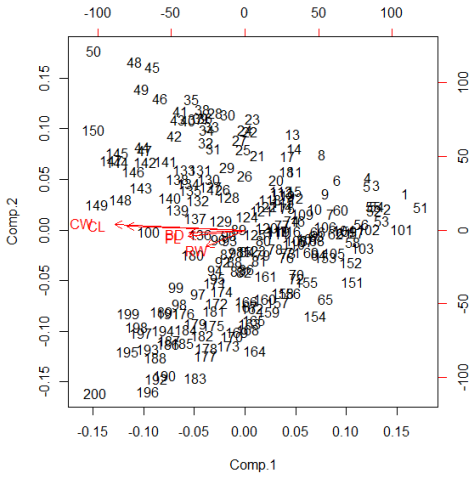

FIGURE 42 – Le plan de représentation 1-2 des individus

Cette ACP confirme bien les résultats obtenus dans la partie 1.2 : les variables sont très corrélées linéairement, à un tel point que le premier axe factoriel dont l'inertie expliquée vaut 98% suffit à lui tout seul à représenter la quasi-totalité de l'information (les variables sont alignées sur l'axe 1).

Solution :
On voit que la variable CW est la plus corrélée à l'axe factoriel 1 et donc à la taille du crabe. On divise donc toutes les autres variables par celle-ci et on l'enlève des données. On refait une analyse sur ces données et on obtient les deux nuages suivants selon le sexe ou l'espèce, où l'on arrive bien a distinguer les individus.

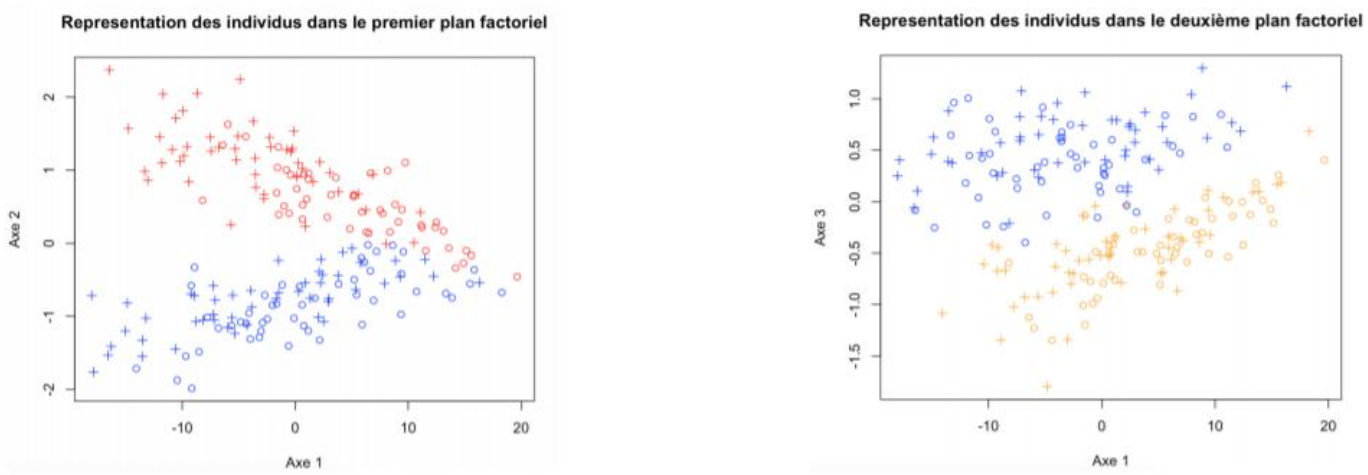

FIGURE 43 – Représentation des individus sur notre jeu de données traité

3.4 Données Pima

On fait une analyse en composantes principales sur le jeu de données Pima :

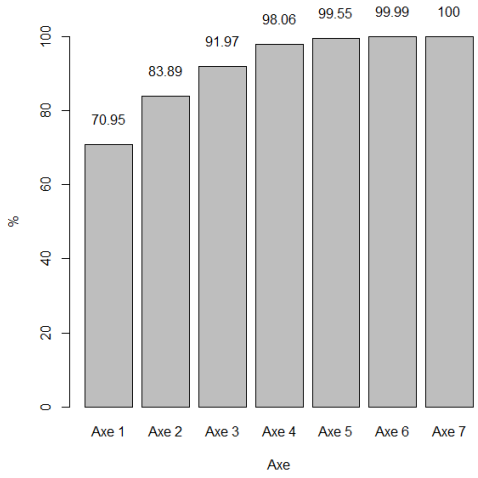

FIGURE 44 – Cumul des inerties expliquées

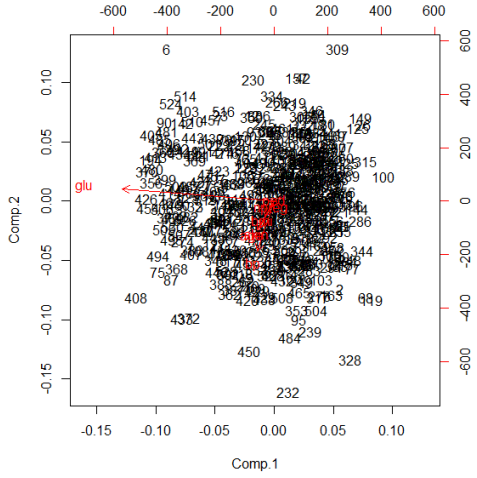

FIGURE 45 – Le plan de représentation 1-2

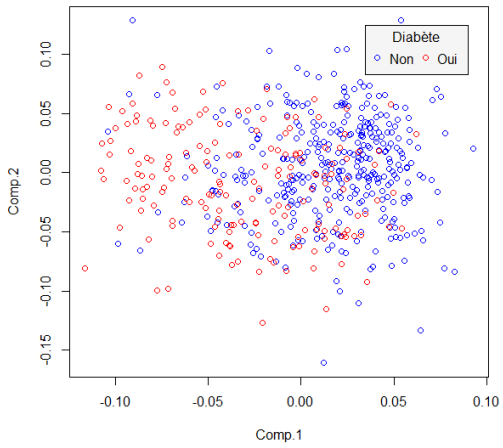

FIGURE 46 – Le plan de représentation 1-2 coloré sur diabète

Sur le plan 1-2, on perde environ 16.11% d'information sur les données ; donc le plan ne traduit pas bien la nature des données. À partir de la Figure 45, on peut déduire que si quelqu'un a une haute valeur de glu, il a une tendance au diabète ; cependant les nuages de 2 populations ne sont pas bien séparés alors on ne peut pas de manière sûre distinguer les deux catégories de patientes.

3.5 Conclusion

Ce premier Travail Pratique nous a permis d'appliquer les méthodes d'analyse de données qui sont enseignées dans l'UV et d'utiliser la méthode ACP très utile pour résumer des informations d'un jeu de données quand le nombre de variables est trop important et que par conséquent on ne peut pas représenter les individus dans un plan.

Nous avons également du utiliser de nombreuses fonctions et commandes de R et par conséquent apprendre et travailler la syntaxe et les possibilités du langage.

Finalement, la mise en place d'un tel document avec les barrières de nos deux langues respectives a nécessité beaucoup d'organisation et de temps pour mener à bien sa rédaction.
En résumé, c'est une expérience très intéressante et formatrice pour notre cursus.

Table des matières

Table des figures