SY09 - TP1 : Statistique descriptive, Analyse en composantes principales

Vançon Virgile et Bathellier Pierre

UTC

Introduction

Ce premier TP de SY09 a pour but de nous familiariser avec les statistiques descriptives et l'analyse en composantes principales (ACP), ainsi que les commandes R. Nous allons pour cela étudier plusieurs jeux de données, comme les résultats des étudiants à l'UV SY02 au printemps 2016, les mesures morphologiques de 200 crabes de sexes différents ou les caractéristiques physiques de 532 femmes dont certaines souffrent de diabète. En appliquant les techniques d'analyse vu en cours, nous allons pouvoir analyser ces tables de la meilleure manière possible et ainsi en extraire le maximum d'informations possibles.

1 Statistique descriptive

1.1 Notes

Description des données

"Notes" est une grande table de 296 entrées (296 étudiants), avec 11 caractéristiques différentes (11 colonnes). La table nous présente un ensemble d'informations relatives aux étudiants ayant suivis SY02 au semestre de printemps 2016.

Description des colonnes :

- nom : variable qualitative nominale. "EtuX" avec X le numéro de l'étudiant.
- specialite : variable qualitative nominale (9 valeurs possible : GB, GI, GM, GP, GSM, GSU, HuTech, ISS et TC), qui correspond au domaine d'étude de l'étudiant.
- niveau : variable quantitative discrète comprise entre 1 et 6, qui précise le semestre de l'étudiant dans sa spécialité.
- statut : variable qualitative nominale, 2 valeurs possible : Echange ou UTC.
- dernier.diplôme.obtenu : variable qualitative nominale (12 valeurs possible : AUTRE 1ER CYCLE, AUTRE 2E CYCLE, AUTRE DIPLOME SUPERIEUR, BAC, BTS, CPGE, DEUG, DUT, ETRANGER SECONDAIRE, ETRANGER SUPERIEUR, INGENIEUR, LICENCE) qui permet de savoir le cursus d'origine de l'étudiant.
- note.median, note.final et note.totale : variables quantitatives continues, comprisent entre 0 et 20.
- correcteur.median et correcteur.final : variable qualitative nominale. "CorX" avec X le numéro du correcteur
- resultat : variable qualitative ordinale (A $<$ B $<$ C $<$ D $<$ E $<$ F $<$ FX).

Analyse des données manquantes

Les valeurs manquantes dans les colonnes note.median, note.final, note.totale, correcteur.median et correcteur.final et resultat correspondent très probablement à des absences aux examens des étudiants concernés, qui n'ont donc ni notes ni correcteurs. On remarque également qu'il manque 6 valeurs dans la colonne dernier.diplome.obtenu, correspondant aux 6 étudiants en échange (l'UTC ne doit pas réclamer les derniers diplômes des étudiants étrangers).

nom	specialite	niveau	statut
0	0	0	0
dernier.diplome.obtenu	note.median	correcteur.median	note.final
6	3	3	12
correcteur.final	note.totale	resultat	
12	12	0	

Figure 1. Nombre de valeurs manquantes par colonne

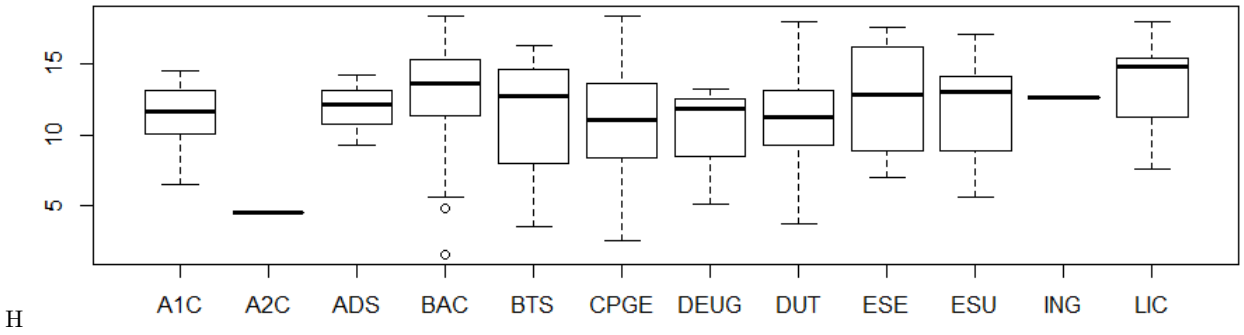
Corrélation entre les variables

Comme on pouvait s'y attendre, on observe une forte corrélation entre les notes de l'examen final et les notes totales et entre les notes de l'examen médian et les notes totales, avec des coefficients de corrélation de 0.92 et de 0,75 (le poids du médian étant moindre). Les notes aux examens conditionnent le résultat de l'UV et le lien entre les notes et la variable résultat est évident.
Il est difficile d'affirmer que d'autres variables sont liées sans faire des tests au préalable.

Influence de la formation d'origine et de la branche sur le résultat obtenu à l'UV

Pour tester l'indépendance entre la branche des étudiants et leurs réussites à l'UV, on réalise un test d'indépendance du χ^2. On calcul tout d'abord la table de contingence, qui donne l'effectif de chaque lettres en fonction des différentes branches. Cependant, pour que le test soit valide, il faut que chaque effectif soit supérieur ou égal à 5, ce qui n'est pas le cas pour toutes les colonnes. Nous avons donc ignoré les étudiants en HuTech, ISS et TC, qui n'étaient pas assez nombreux.
Après réalisation du test, on obtient une p-value = 0.5107, et on accepte donc l'hypothèse d'indépendance entre la formation d'origine et le résultat obtenu à l'UV, pour un test à 95% de confiance.

Pour tester l'indépendance entre le dernier diplôme obtenu et le résultat à l'UV, on réalise un autre test d'indépendance du χ^2. Comme pour le test précédent, il est préférable de ne pas tenir compte de la majorité des colonnes ayant des effectifs trop faibles. Les 3 seuls formations d'origine ayant des effectifs assez importants sont le BAC (106), le DUT (89) et les CPGE (39).
On obtient alors une p-value très faible de 0.0006389, qui rejette l'hypothèse d'indépendance, pour un test à 95% de confiance. On peut voir sur la figure 2 ci-dessous que les étudiants ayant intégré l'UTC post-bac réussissent mieux que ceux venant des CPGE ou de DUT.
Si on prend en compte toutes les colonnes, on obtient une p-value de 0.25, qui aurait confirmé l'hypothèse d'indépendance. Il est donc important de bien respecter les conditions nécessaires à la réalisation du test.



H

Figure 2. Boîte à moustache des notes totales en fonction du dernier diplôme obtenu

Les boîtes à moustache ci-dessus montrent que les étudiants dont le dernier diplôme est le bac ou une licence réussissent bien mieux que ceux issus des CPGE. Ces derniers constituent un groupe très hétérogènes, avec de très bons élèves et d'autres en grande difficulté, contrairement à ceux ayant une licence par exemple.

Influence du correcteur sur les notes du médian et du final

Le test d'indépendance du χ^2 n'est pas réalisable sur la table de contingence des notes du médian en fonction du correcteur, car les effectifs pour chaque note sont bien trop faibles. Nous avons donc regroupé les colonnes en 4 plages de données afin d'augmenter les effectifs : [0,5[, [5,10[, [10,15[, [15,20]. On passe alors d'une matrice 7x37 à une matrice 7x4.
On obtient une p-value de 0,6523 : les correcteurs n'ont a priori pas d'influence sur les notes du médian, pour un test à 95% de confiance.

Les boîtes à moustache ci-dessus ont tendances à confirmer l'indépendance entre le correcteur et les notes obtenues au médian. En effet, les médianes sont sensiblement les mêmes et seuls les écarts inter-quartiles évoluent significativement, notamment pour le correcteur 5.

De même pour l'indépendance entre la note au final et le correcteur, il faut rassembler les données de la table de contingence, selon les mêmes plages. On passe alors d'une matrice 7x36 à une matrice 7x4.
On obtient une p-value de 0.043, qui rejetterait l'hypothèse d'indépendance entre le correcteur et la note obtenue au final, pour un test à 95% de confiance. Avec un test à 99% de confiance, notre conclusion aurait été différente. Les notes au final semblent dans tout les cas plus liées au correcteur que les notes du médian.

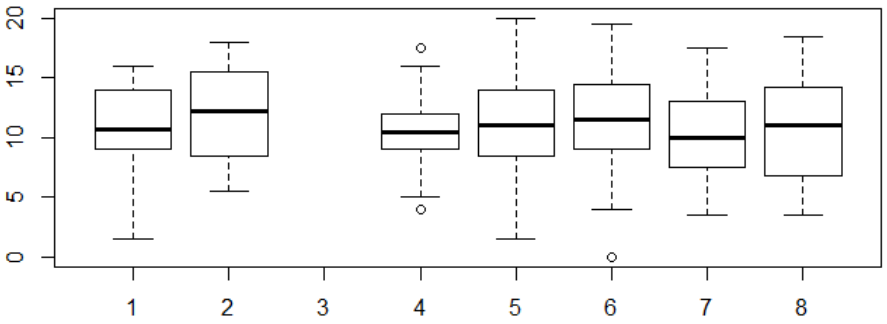


Figure 3. Boîte à moustache des notes du median en fonction des correcteurs

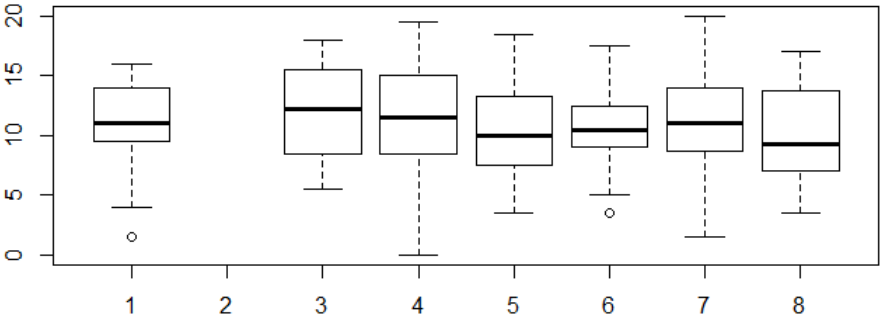


Figure 4. Boîte à moustache des notes du final en fonction des correcteurs

1.2 Données Crabs

Description des données

Dans cette partie nous allons analyser les données "Crabs". Crabs est un jeu de données composé de 200 observations et de huit variables. Les données décrivent les mesures morphologiques de 200 crabes des deux sexes et de deux couleurs différentes. Les 200 individus sont répartis en 4 groupes, chacun ayant un effectif de 50, ayant un sexe et étant d'une espèce particulière. (50 crabes orange/féminin, 50 bleu/masculin etc ...). Il n'y a pas de données manquantes dans la table.

Description des colonnes :

- sp : Variable qualitative nominale. Renseigne l'espèce de l'individu. Valeurs possibles : ["O"/"B"]
- sex : Variable qualitative nominale. Renseigne le sexe de l'individu. ["M"/"F"]
- index : Variable qualitative ordinale. Numéro de l'individu au sein de son groupe de 50 crabes particulier. [1-50]
- FL : Variable quantitative continue. Taille du lobe frontal en mm.
- RW : Variable quantitative continue. Largeur arrière en mm.
- CL : Variable quantitative continue. Longueur de la carapace en mm.
- CW : Variable quantitative continue. Largeur de la carapace en mm.
- BD : Variable quantitative continue. Profondeur du corps en mm.

Les ordres de grandeur des caractéristiques morphologiques sont globalement dans le même ordre de grandeur et vont de 6,5 mm à 54,60mm.

Recherche de différences morphologiques en fonction du sexe

Quand on compare les répartitions des caractéristiques morphologiques en fonction du sexe, seule une différence au niveau de la largeur arrière (RW) semble vraiment marquante. A première vu, le sexe ne semble pas avoir d'influence sur les caractéristiques morphologiques des crabes.

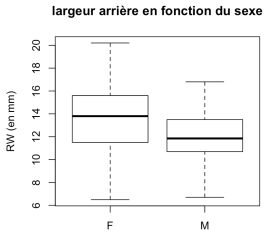


Figure 5. Largeur arrière en fonction du sexe

En revanche, les différences sont davantage marquées au niveau des espèces. En effet, il y a une réelle différence entre la répartition des tailles des fronts (FL), des largeurs arrières (RW), des longueurs de carapace (CL) et des profondeurs du corps.

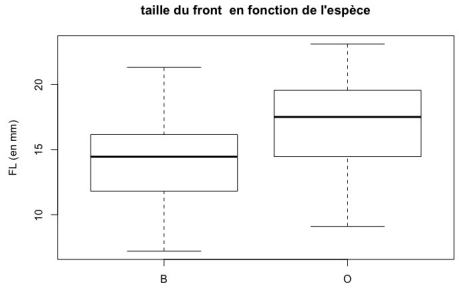


Figure 6. Taille du front en fonction de l'espèce

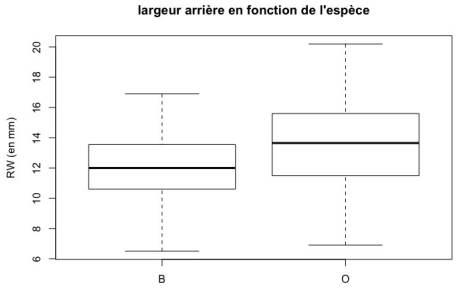


Figure 7. Largeur arrière en fonction de l'espèce

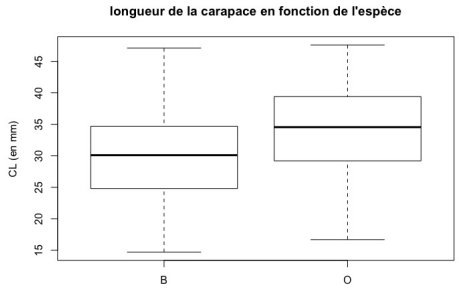


Figure 8. Longueur de la carapace en fonction de l'espèce

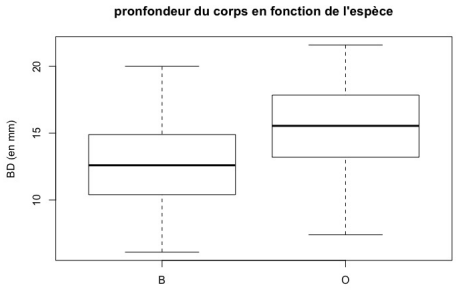


Figure 9. Profondeur de l'espèce

Il faut donc chercher un autre moyen de pouvoir identifier le sexe à partir des caractéristiques morphologiques. Si on réalise la matrice des corrélations et que l'on distingue chaque point par le sexe et une autre fois par l'espèce on obtient les graphes suivant :

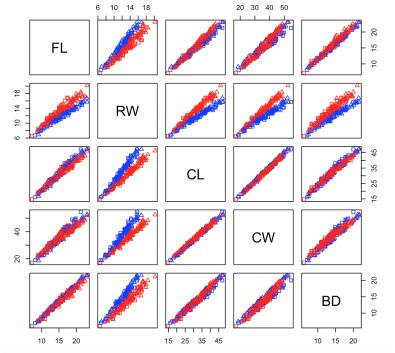


Figure 10. Graphe des corrélations - Distinction par sexe (F=red,M=Blue)

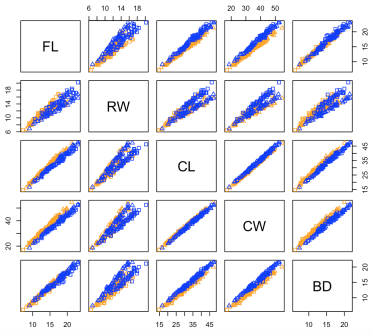


Figure 11. Graphe des corrélations - Distinction par espèce (O=orange,B=Blue)

Au contraire de la distinction par espèce où il semble compliqué de dégager une tendance, certaines paires de caractéristiques morphologiques font apparaître une distinction en fonction du sexe. On s'intéresse par exemple à la paire RW/CL :

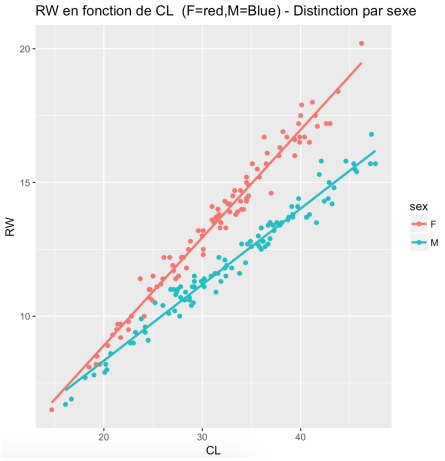


Figure 12. RW en fonction de CL (F=red,M=Blue)

Si l'on trace deux courbes de tendance en regroupant les individus par sexe on obtient deux droites de coefficient directeur :

— 0.4 dans le cas des individus de sexe féminin.
— 0,25 dans le cas des individus de sexe masculin.

Il semble ainsi possible de déterminer le sexe de l'individu en calculant le rapport entre certaines de ses caractéristiques notamment sa largeur arrière sur sa longueur de carapace.

Corrélation entre les variables

Le graphe des corrélations (Figure 10 et 11) témoigne d'une forte corrélation linéaire entre les variables. La matrice des corrélations présentée ci-dessous tend à confirmer ce résultat, les corrélations entre les différentes variables étant systématique supérieures ou égales à 89% :

	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 13. Matrice des corrélations caractéristiques physiques des crabes

Ces corrélations paraissent plutôt logiques, le jeu de données étant des caractéristiques physiques. Si l'on prend l'exemple de l'Homme, plus un individu est grand plus on s'attend à ce que ses caractéristiques physiques (longueur des jambes, longueur du torse etc...) soient grandes. Dans ce jeu de données, une caractéristique physique importante témoigne certainement d'un crabe assez gros. Il paraît alors logique que le reste de ses caractéristiques physiques soient elles aussi importantes.

Un moyen de s'affranchir de cette corrélation pourrait être de regrouper les variables entre elles, ou même de limiter leur nombre. En effet, accumuler autant de variables aussi liées n'apporte pas réellement d'informations.

1.3 Pima

Description des données

"Pima" est une grande table de 532 entrées (532 femmes), caractérisées par 8 colonnes. Cette table nous présente un ensemble de caractéristiques physiques et génétiques chez les femmes, dans le but de comprendre les causes probables d'un diabète.

Description des colonnes :

— npreg : variable quantitative discrète à valeurs dans [0, 17]. Indique le nombre de grossesses.
— glu : variable quantitative discrète à valeurs [56, 199]. Indique le taux plasmatique de glucose.
— bp : variable quantitative discrète à valeurs dans [24, 110]. Indique la pression artérielle diastolique.
— ski : variable quantitative discrète à valeurs dans [7, 99]. Indique l'épaisseur du pli cutané au niveau du triceps.
— bmi : variable quantitative continue, comprise entre 18.2 et 67.1 pour notre jeu de données. Indique l'indice de masse corporelle.

— ped : variable quantitative continue, comprise entre 0.085 et 2.42 pour notre jeu de données. Indique la fonction de pedigree du diabète.
— age : variable quantitative discrète, à valeurs dans [21, 81].
— z : variable qualitative nominale, avec 2 valeurs possible : 1 si l'individu n'est pas diabétique ou 2 s'il l'est.

Description des données

Il n'y a aucune valeur manquante dans la table Pima, ce qui évitera d'avoir à faire des traitements supplémentaires sur les données.

Étude des liens statistiques entre les variables

Afin d'étudier les liens statistiques entre les variables de la matrice Pima nous utiliserons deux fonctions R permettant de visualiser les résultats d'une matrice de corrélation :

— *corrplot* : cette fonction permet de visualiser graphiquement le résultat d'une matrice de corrélation
— *symnum* : cette fonction remplace les coefficients de corrélation par des symboles en fonction de la valeur. Elle prend la matrice de corrélation comme argument. Nous avons ici décidé de représenter tout les liens forts (ceux pour lesquels la corrélation est supérieure à 0,5) par des "+".

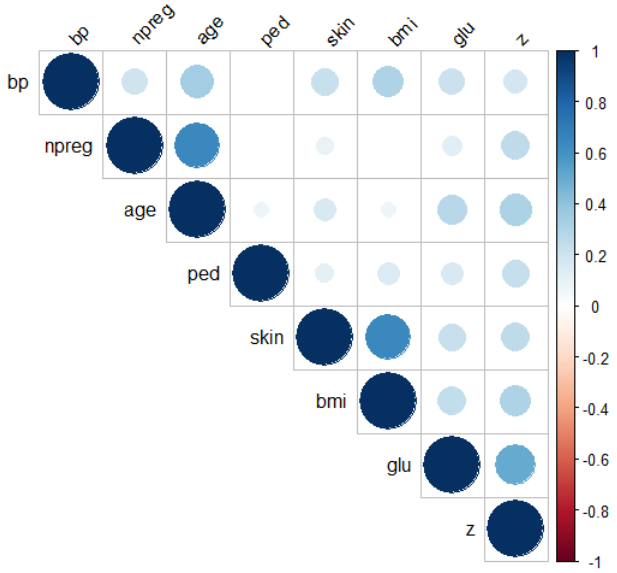


Figure 14. Corrplot de la matrice de corrélation

```
> symnum(cor(Pima2), cutpoints = c(0, 0.5, 0.95, 1), symbols = c('', 'x', '/'))
      n g bp s bm p a z
npreg /             x
glu     /             x
bp        /
skin        / x
bmi         x /
ped               /
age   x             /
z       x             /
attr(,"legend")
[1] 0 '' 0.5 'x' 0.95 '/' 1
```

Figure 15. symnum de la matrice de corrélation (corrélograme)

La première figure nous permet d'observer tout les liens de corrélation existant entre les variables, et la deuxième figure ne fait apparaître que les liens forts (représentés par des "+"). On observe alors 3 liens statistiques forts :

— entre npreg (nombre de grossesses) et age : ce lien est logique, car plus une femme est âgée et plus elle a eu la possibilité d'avoir des enfants.
— entre skin (épaisseur du pli cutané au niveau du triceps) et bmi (indice de masse corporelle) : ce lien est également évident, car l'épaisseur du pli cutanée est un critère de calcul de la masse grasse chez un individu, et donc de l'IBM.

— entre z (indique si l'individu en question est diabétique) et glu (taux plasmatique de glucose).
Le diabète sucré est une maladie liée à une défaillance des mécanismes biologiques de régulation de la glycémie (concentration de glucose dans le sang) menant à une hyperglycémie. La concentration de glucose dans le sang est ici représentée par le taux plasmatique de glucose, et la corrélation est donc "normale" entre les deux variables.
Il est alors intéressant de noter qu'il existe deux types de diabète : le diabète de type 1 qui est une maladie auto-immune et qui représente environ 6% des cas, et le diabète de type 2 qui apparaît essentiellement chez les individus de plus de 40 ans en sur-poids, et qui représente environ 90% des cas. L'absence de corrélation entre le bmi et z et entre age et z nous permet d'affirmer que les femmes diabétique de notre échantillon souffre très probablement d'un diabète de type 1 (type insulino-dépendant).

2 Analyse en composantes principales

2.1 Exercice théorique

Notre jeu de données contient les moyennes et les écarts-types par correcteur, pour le médian et le final de la table notes présentée dans la première partie du TP.
Pour déterminer les axes factoriels du nuage de points on procède de la manière suivante :

— On centre la matrice
— On détermine la matrice de covariance en appliquant la formule

$$V = X^T DpX \tag{1}$$

On obtient les axes factoriels suivant en diagonalisant la matrice de covariance :

	1	2	3	4
1	-0.036	-0.704	-0.233	0.669
2	0.294	-0.647	-0.094	0.697
3	-0.955	-0.172	-0.021	0.241
4	-0.001	-0.236	0.968	0.088

Figure 16. Axes factoriels de la matrice des correcteurs

On détermine le pourcentage d'inertie de chacun de ces axes.

	U1	U2	U3	U4
Inertie explique	66.095613	24.786460	5.610544	3.507382
Inertie cumule	66.095613	90.882073	96.493517	100

Figure 17. Pourcentage d'inertie expliqué de chacun des axes et inertie cumulée

On remarque que l'inertie cumulée des deux premiers axes factoriels U1 et U2 dépasse 90%. On en conclue donc que ces deux premiers axes expliquent très bien les informations du nuage de points.

On peut, grâce aux vecteurs propres, déterminer la matrice des composantes principales en appliquant la formule :

$$C = XMU \tag{2}$$

Cela nous permet d'obtenir la matrice des composantes principales, on peut ensuite faire une représentation des individus dans le premier plan factoriel en isolant les deux premières colonnes de la matrice des composantes principales.

	[,1]	[,2]	[,3]	[,4]
1	1.1131294	0.2973223	0.10675046	0.40247613
4	-1.5041532	0.8133820	0.01415599	0.06168357
5	0.4045437	-0.2338454	-0.61494553	-0.04153965
6	-1.1613043	-1.0159209	0.11740385	0.08203415
7	0.2520651	0.4929044	0.07733933	-0.32451159
8	0.8957193	-0.3538424	0.29929590	-0.18014261

Figure 18. Matrice des composantes principales

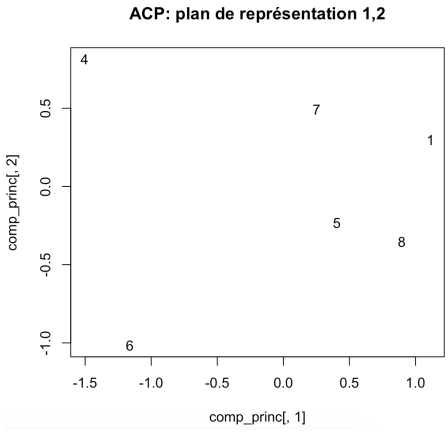


Figure 19. Représentation des six individus dans le premier plan factoriel

On peut maintenant, grâce au calcul des corrélations, faire une représentation des quatre variables dans le premier plan factoriel :

$$Cor(\alpha, j) = diag(\frac{1}{\sqrt{\frac{N-1}{N}} * apply(X, 2, sd)}) * U * \sqrt{L} \tag{3}$$

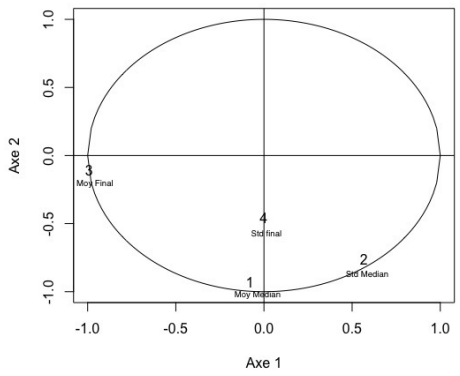


Figure 20. Représentation des quatre variables dans le premier plan factoriel

On en conclue que l'axe 1 est principalement corrélé avec la moyenne au final et dans une moindre mesure à l'écart type au médian. L'axe 2 est quant à lui fortement corrélé à la moyenne au médian et à moindre mesure à l'écart type au final.

Recomposition :
Calcul de l'expression $\sum_{\alpha=1}^{k} c_\alpha u_\alpha^T$: pour k = 4 on obtient la matrice :

	[,1]	[,2]	[,3]	[,4]
[1,]	1.1131294	0.2973223	0.10675046	0.40247613
[2,]	-1.5041532	0.8133820	0.01415599	0.06168357
[3,]	0.4045437	-0.2338454	-0.61494553	-0.04153965
[4,]	-1.1613043	-1.0159209	0.11740385	0.08203415
[5,]	0.2520651	0.4929044	0.07733933	-0.32451159
[6,]	0.8957193	-0.3538424	0.29929590	-0.18014261

Figure 21. Matrice initial centrée

On remarque, pour k = 4, que l'on obtient la matrice issue de corr.acp après centrage. Cette opération dites de "reconstitution" revient à reconstituer une matrice à partir de ses composantes principales et des axes principaux, c'est à dire ses vecteurs propres.

Imputations par la moyenne :
On repart de jeu de données initial en corrigeant les valeur NA par la moyenne de la variable correspondante. 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 22. Jeu initial corrigé avec la moyenne

La méthode ACP détaillé précédemment est ré-appliquée avec ces nouvelles données. Cela nous permet d'obtenir une nouvelle représentation des huit correcteurs dans les deux premiers plans factoriels :

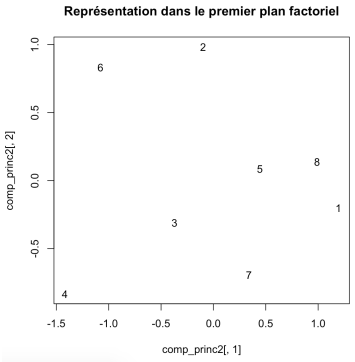


Figure 23. Représentation des huit correcteurs dans le premier plan factoriel

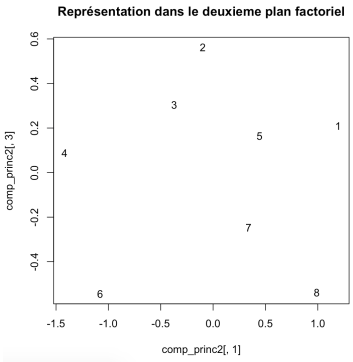


Figure 24. Représentation des huit correcteurs dans le deuxième plan factoriel

On retrouve la répartition des individus des correcteurs de la figure 23. Il est normal que les deux correcteurs ajoutés ne bouleversent pas la représentation dans le premier plan factoriel. En effet les valeurs que l'on a prise pour ces individus étant des moyennes ils se retrouveront centrés dans les nuages de point de représentation des individus.

2.2 Utilisation des outils R

Pour réaliser notre ACP on a recourt à la fonction *princomp*. Les résultats que l'on a obtenu précédemment sont récupérables grâce à des accesseurs sur la variable retournée par *princomp*.

— **Fonction summary** : Cette fonction nous permet de retrouver l'inertie expliquée de chaque axe et l'inertie cumulée. La ligne "Proportion of Variance" correspond à l'inertie expliquée. La ligne "Cumulative Proportion" correspond à l'inertie cumulée.

```
> summary(acp)
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4
Standard deviation     0.9899215 0.6062080 0.28841438 0.22803734
Proportion of Variance 0.6609561 0.2478646 0.05610544 0.03507382
Cumulative Proportion  0.6609561 0.9088207 0.96492618 1.00000000
```

Figure 25. Fonction summary sur l'acp

— **Fonction acp$scores** : Cette fonction nous permet de retrouver les composantes principales calculées précédemment. On peut facilement retrouver les valeurs propres grâce à la relation $\lambda_\alpha = \frac{\sum c_{i\alpha}^2}{n}$

```
> acp$scores
      Comp.1     Comp.2     Comp.3     Comp.4
1  1.1131294  0.2973223  0.10675046  0.40247613
4 -1.5041532  0.8133820  0.01415599  0.06168357
5  0.4045437 -0.2338454 -0.61494553 -0.04153965
6 -1.1613043 -1.0159209  0.11740385  0.08203415
7  0.2520651  0.4929044  0.07733933 -0.32451159
8  0.8957193 -0.3538424  0.29929590 -0.18014261
```

Figure 26. Fonction scores sur l'acp, récupération des vecteurs propres

```
> #calcul des valeurs propres
>   valeurspropres = list()
>   for(i in seq(1,4))
+   {
+     Lambda = 0
+   for (alpha in seq(1, 6)) {
+     Lambda = (Lambda + acp$scores[alpha,i]^2)
+   }
+     Lambda=Lambda/6
+   valeurspropres <- cbind(valeurspropres, Lambda)
+   }
> valeurspropres
     Lambda     Lambda     Lambda     Lambda
[1,] 0.9799445 0.3674882 0.08318286 0.05200103
```

Figure 27. Calcul des valeurs propre à partir des vecteurs propres

— **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. 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 peut être s'expliqué par le fait que notre ACP a été réalisée biaisée dans l'exercice 2.1.

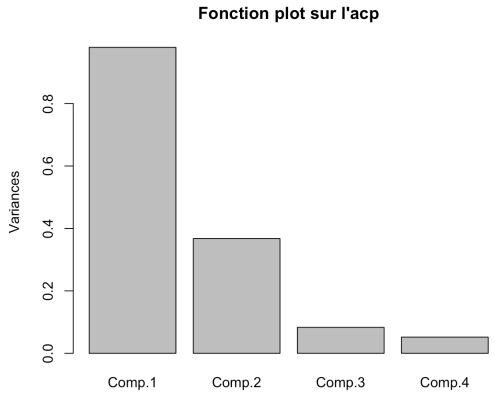


Figure 28. Fonction plot sur l'acp

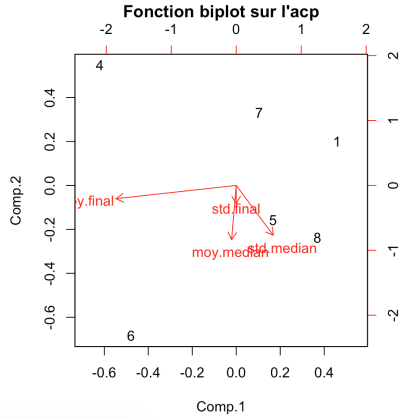


Figure 29. Fonction biplot sur l'acp

2.3 Données Crabs

On réalise une première ACP sur Crabsquant. On visualise les résultat grâce à biplot :

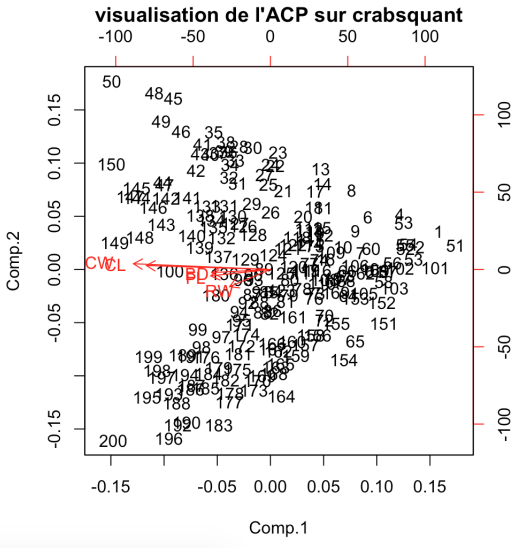


Figure 30. Représentation dans le premier plan factoriel des individus et des variables

L'ACP vient confirmer l'analyse de la partie 1.2. Les variables sont très corrélées linéairement si bien que le premier axe factoriel, dont l'inertie expliquée vaut 98%, suffit à représenter la quasi totalité de l'information (les variables sont quasiment alignées sur l'axe 1).

Tentative de limitation du facteur de taille du crabe

D'après la première ACP sur le jeu de données, la variable CW est la variable qui est la plus fortement corrélée à l'axe 1 et donc à la taille du crabe. On traite donc le jeu de données en sortant cette variable du tableau et en divisant le reste des caractéristiques par cette variable. Lorsque l'on refait l'ACP sur ce jeu de données et que l'on représente les des premiers plans factoriel on constate que l'on arrive facilement à distinguer les individus :

— Par le sexe dans le premier plan factoriel
— Par l'espèce dans le second plan factoriel

.

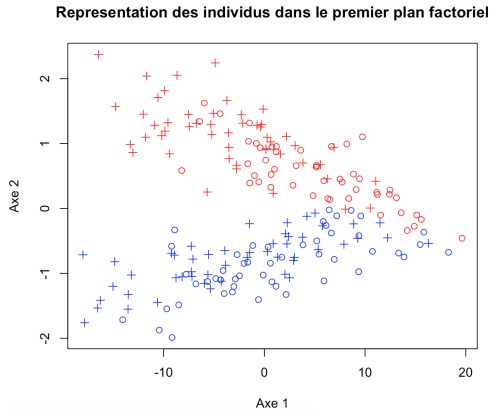


Figure 31. Jeu de données traité - Représentation des individus dans le premier plan factoriel (Red="F", Blue="M")

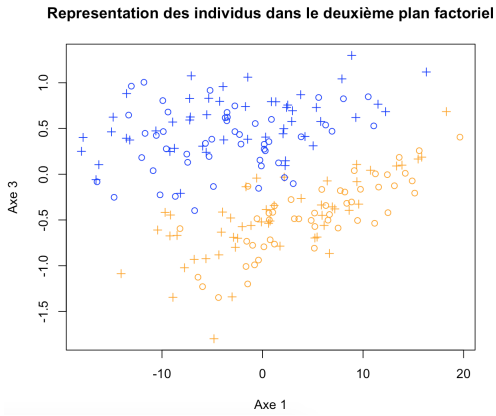


Figure 32. Jeu de données traité - Représentation des individus dans le second plan factoriel (Orange="O", Blue="B")

2.4 Données Pima

L'utilisation de la fonction *princomp* sur les données Pima engendre une erreur R.

```
> acp = princomp(Pima)
Error in cov.wt(z) : 'x' ne doit contenir que des valeurs définies
```

Figure 33. Error R après utilisation de la fonction *princomp* sur Pima

Cette erreur est due à l'utilisation de la fonction factor sur la dernière colonne de Pima au moment de son initialisation, i.e. la colonne indiquant si une femme est diabétique, afin de transformer cette variable quantitative discrète en variable qualitative nominale. Or une ACP ne peut s'effectuer que sur des variables quantitatives.
Il est possible de contourner ce problème en évitant d'utiliser la fonction factor, et en faisant une ACP «manuelle» comme dans l'exercice 2.1, i.e. en évitant d'utiliser la fonction *princomp*.
Le calcul des corrélations $\mathrm{Cor}(\alpha, j)$ (équation (3)) nous permet de voir que les colonnes 2 et 8 de la table Pima, i.e. le taux plasmatique de glucose et le fait que l'individu en question soit diabétique sont fortement liées.

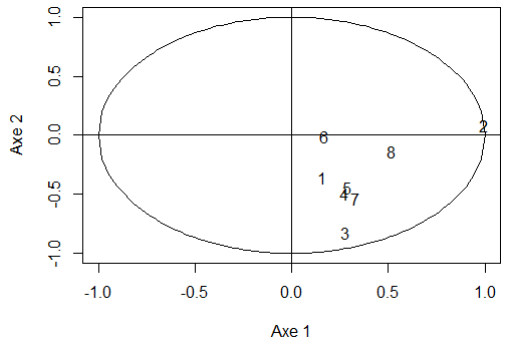


Figure 34. Visualisation du calcul des corrélations Cor(α, j) des données Pima

Cette visualisation montre que l'axe 1 calculé par l'ACP explique très bien le taux plasmatique de glucose et le fait que l'individu en question soit diabétique ou non. Ceci confirme notre précédente conclusion affirmant que ces 2 variables sont corrélées.

Il semble compliqué de trouver une représentation simple permettant de séparé les deux catégories de variables. Cependant, l'ACP réalisée peut nous donner une idée, le premier axe expliquant fortement le taux plasmatique de glucose et donc le fait qu'une femme soit diabétique.

Conclusion

Nous avons donc pu utiliser, tout au long de ce TP, les méthodes d'analyse enseignées en SY09 et particulièrement l'analyse en composantes principales (ACP). Nous avons également mis en place un large panel de commandes et d'outils R. Cette mise en pratique nous a permis de nous rendre compte qu'il était impossible d'appliquer simplement les formules vu en cours, mais qu'une analyse personnelle poussée était nécessaire dans chaque cas afin de répondre au mieux aux questions soulevées par le sujet.

Table des figures