SY09 - Analyse des données et Data-Mining

Printemps 2017 (P17)

TP1

Statistique descriptive, Analyse en composantes principales

Maxime Churin GI04 - Louis Denoyelle GI04

Introduction

Ce TP de SY09 s'inscrit dans la phase exploratoire de l'UV. La première partie consiste à utiliser la statistique descriptive pour vérifier la cohérence de nos jeux de données et la recherche de relations entre les variables. La seconde partie va la compléter par une analyse en composantes principales dans le but d'améliorer la visualisation en réduisant le nombre de variables tout en perdant le moins d'informations. Nous nous appuierons sur les outils de représentation et de calcul du logiciel R pour analyser ces trois jeux de données.

1 Données notes

1.1 Statistique descriptive

Dans cette première partie, nous allons analyser de façon descriptive le jeu de données sy02-p2016.csv contenant les 296 étudiants inscrits à l'UV SY02 au semestre de printemps 2016 décrits par un ensemble de 11 variables.

On peut distinguer ces variables en trois groupes. Le premier décrit les variables qualitatives nominales, il est composé de 6 variables : *nom, specialite, statut* (université d'origine), *dernier diplome obtenu* et *correcteur median/final*. Le second groupe comprend les deux variables qualitatives ordinales que sont *niveau* (indice du semestre) et *resultat* (notation via la norme ECTS). Pour terminer on retrouve le groupe des variables quantitatives contenant *note median/final* avec une précision de 0.5 et la variable synthétique résultante *note totale* avec une précision de 0.1.

Par ailleurs, on observe l'absence de certaines données. C'est le cas de la variable *dernier diplome obtenu* lorsqu'elle concerne des étudiants étrangers car on ne possède tout simplement pas cette information. C'est aussi valable pour le couple *note median/final*, le couple *correcteur median/final* et la variable résultante *note totale*. En effet, lorsqu'un étudiant ne s'est pas présenté au médian ou au final, il est impossible de lui attribuer une note et un correcteur.

Après une rapide analyse visuelle, nous sommes tentés de supposer la dépendance entre certaines variables. Les études en France et particulièrement à l'UTC, étant assez réputées, les étudiants en échange devraient obtenir de moins bons résultats. Enfin, même si cela est évident, des notes découlent des variables synthétiques, elles sont donc étroitement liées avec *note totale* et *resultat*.

Nous allons maintenant étudier l'impact de différentes variables sur la réussite de l'UV. Pour statuer sur ces liens statistiques, nous avons décidé d'utiliser *note totale* plutôt que *resultat*. En effet, pour compléter les observations graphiques nous avons réalisé des tests statistiques et ceux-ci sont plus pertinents avec une variable continue que discrète. La condition de découpage du $\tilde{\chi}^2$ n'étant jamais atteinte, nous réalisons des tests d'analyse de la variance. Nous avons choisi de représenter les données sous formes de diagrammes à moustache incluant la notion d'effectif représentée par la taille de la boîte. C'est la représentation que nous avons trouvé la plus adaptée à ces données.

Nous avons effectué un pré-traitement afin d'enlever la modalité HuTech de la variable *specialite* car elle n'était pas représentative. Malgré les effectifs différents nous ne pouvons pas envisager de lien entre ces deux variables.

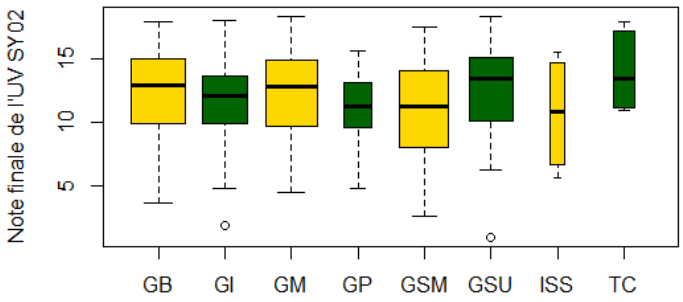


Figure 1.1 – Influence de la specialité sur la note totale.

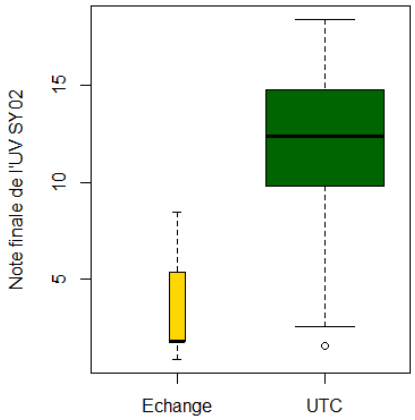


Figure 1.2 – Influence de la formation d'origine sur la note totale.

On remarque tout de suite l'impact des formations françaises. En effet, aucun étudiant étranger n'arrive à obtenir la moyenne et ne peut valider l'UV. D'après ce *boxplot*, nous pouvons déduire que la variable *statut* est directement liée à la réussite de l'UV. Nous allons le prouver avec une analyse de la variance. La p-value résultante de ce test étant très basse et inférieure au seuil de confiance 1%, on rejette l'hypothèse H_0. Nous concluons donc que les modalités de la variable *statut* ont une influence significative sur la variable *note totale* : les moyennes des étudiants étrangers et UTC sont significativement différentes.

On observe ici des variables à priori liées ce qui signifierait que l'indice du semestre influerait sur la variable *note totale*. Il en découle que plus SY02 est faite tôt meilleur sont les résultats. C'est plutôt surprenant, car nous ne nous attendions pas à tirer de telles conclusions de cette étude. La réalisation d'un test ANOVA, confirme cette hypothèse au seuil de confiance 1%, à savoir que les modalités de la variable *niveau* ont une influence significative sur la variable *note totale*.

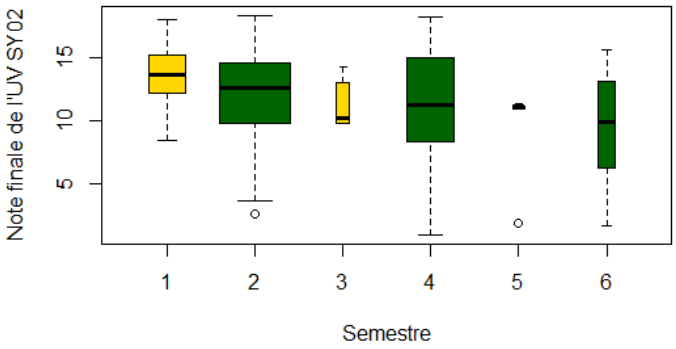


Figure 1.3 – Influence du semestre sur la note totale

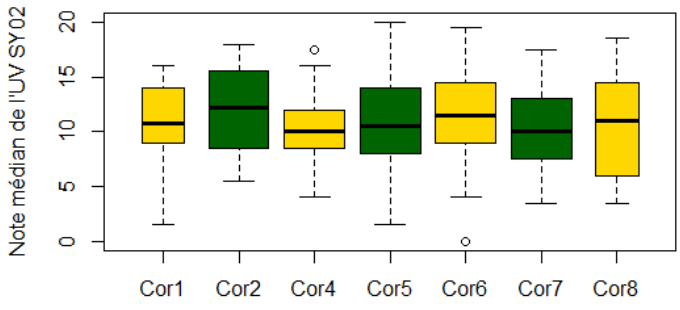


Figure 1.4 – Influence du correcteur sur la note du médian.

Ce graphique ne nous permet pas d'envisager un quelconque lien entre ces deux variables. On observe cependant une médiane qui coupe en deux parties plus ou moins égales l'espace interquartiles. Cette observation pourrait se justifier par les consignes que reçoivent les correcteurs pour l'uniformisation des notes malgré des effectifs d'élèves différents.

1.2 Analyse en composantes principales

1.2.1 Exercice théorique

On décide maintenant d'utiliser une méthode d'analyse appelée ACP sur les correcteurs du jeu de données. On construit donc la matrice qui agrège les correcteurs du médian et du final décrits par leur moyenne et leur écart type, à cause de l'absence de certains correcteurs pour le médian ou le final nous les excluons de l'étude. On obtient alors la matrice X composée de six individus et quatre variables. La même pondération étant appliquée sur chaque individu et l'ensemble muni de la métrique euclidienne, on définit les matrices $M = I_5$ et $D_6 = \frac{1}{6} I_6$. On commence par centrer la matrice X en soustrayant sa moyenne à chaque colonne (utilisation de

la fonction `scale` de R) :

$$X = \begin{pmatrix} 10.71 & 3.90 & 10.94 & 4.58 \\ 10.23 & 3.04 & 13.43 & 4.34 \\ 10.98 & 4.41 & 11.83 & 3.97 \\ 11.50 & 4.30 & 13.41 & 4.88 \\ 10.12 & 4.03 & 11.90 & 4.44 \\ 10.74 & 4.65 & 11.40 & 4.87 \end{pmatrix} \qquad \overline{X} = \begin{pmatrix} -0.006 & -0.156 & -1.213 & 0.068 \\ -0.479 & -1.013 & 1.282 & -0.172 \\ 0.265 & 0.357 & -0.323 & -0.544 \\ 0.786 & 0.247 & 1.262 & 0.362 \\ -0.592 & -0.026 & -0.249 & -0.071 \\ 0.026 & 0.590 & -0.757 & 0.357 \end{pmatrix}$$

On calcule ensuite la matrice de variance V selon la formule suivante :

$$V = \overline{X}^T D_6 \overline{X} = \begin{pmatrix} 0.211 & 0.134 & 0.071 & 0.046 \\ 0.134 & 0.265 & -0.226 & 0.045 \\ 0.071 & -0.226 & 0.908 & 0.013 \\ 0.046 & 0.045 & 0.013 & 0.099 \end{pmatrix}$$

Puis nous diagonalisons cette matrice V pour obtenir les vecteurs propres qui correspondent aux axes factoriels et leurs valeurs propres associées triées par ordre décroissant (la fonction `eigen` de R). On obtient $\lambda_1 = 0.98, \lambda_2 = 0.37, \lambda_3 = 0.08$ et $\lambda_4 = 0.05$ avec

$$U_1 = \begin{pmatrix} -0.037 \\ 0.294 \\ -0.955 \\ -0.001 \end{pmatrix}, U_2 = \begin{pmatrix} -0.704 \\ -0.647 \\ -0.172 \\ -0.236 \end{pmatrix}, U_3 = \begin{pmatrix} -0.233 \\ -0.094 \\ -0.021 \\ 0.968 \end{pmatrix} U_4 = \begin{pmatrix} 0.669 \\ -0.697 \\ -0.241 \\ 0.088 \end{pmatrix}$$

On peut alors calculer les pourcentages d'inertie expliquée par chacun de ces axes selon la formule : $E_k = \frac{\lambda_k}{\sum_{i=1}^{4} \lambda_i}$, ce qui nous donne : $E_1 = 66\%, E_2 = 25\%, E_3 = 5.8\%$ et $E_4 = 3.6\%$.

Pour terminer, on calcule la matrice des composantes principales avec U la matrice des vecteurs propres :

$$C = \overline{X} M U = \begin{pmatrix} 1.11 & 0.30 & 0.11 & 0.40 \\ -1.50 & 0.81 & 0.01 & 0.06 \\ 0.40 & -0.23 & -0.61 & -0.04 \\ -1.16 & -1.02 & 0.12 & 0.08 \\ 0.25 & 0.49 & 0.08 & -0.32 \\ 0.90 & -0.35 & 0.30 & -0.18 \end{pmatrix}$$

A partir de cette matrice, on peut, par exemple, tracer la représentation des six individus dans le premier plan factoriel. Une autre représentation intéressante est celle des variables initiales dans le premier plan factoriel.

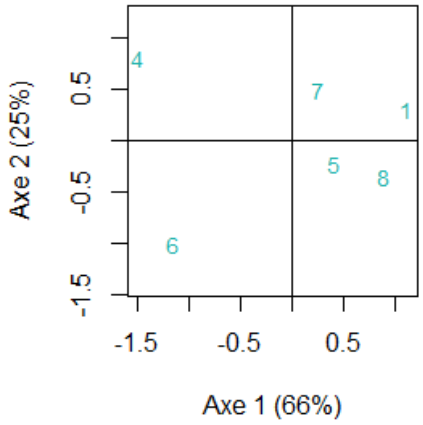


Figure 1.5 – ACP sur les individus (correcteurs).

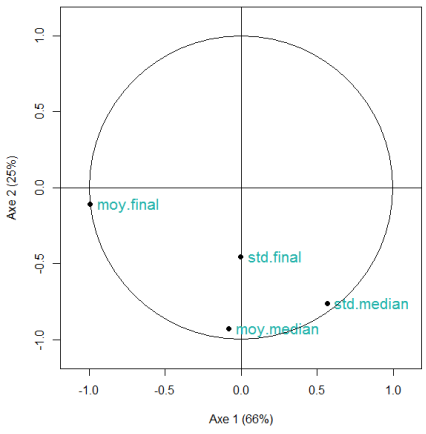


Figure 1.6 – ACP sur les variables.

Cette dernière permet de visualiser les liens entre les variables initiales mais aussi les liens entre les composantes principales et les variables initiales. Pour l'obtenir nous devons calculer les corrélations entre les variables selon la formule :

$$Cor(\alpha, j) = \frac{1}{\sqrt{\overline{x}_{jj}}} * \frac{1}{\sqrt{\lambda_\alpha}} * \overline{x}_j^T * D_p * c_\alpha \quad Cor = \begin{pmatrix} -0.079 & -0.928 & -0.146 & 0.332 \\ 0.566 & -0.762 & -0.053 & -0.309 \\ -0.992 & -0.109 & -0.006 & -0.058 \\ -0.002 & -0.456 & 0.888 & 0.064 \end{pmatrix}$$

On constate que dans le premier plan factoriel les variables *moy.median, std.median* et *moy.final* sont bien représentées car elles sont plus éloignées du centre que la variable *std.final*. En utilisant la formule $\sum_{\alpha=1}^{k} c_\alpha u_\alpha^T$, on obtient :

$$k = 1 : \begin{pmatrix} -0.041 & 0.327 & -1.063 & -0.001 \\ 0.055 & -0.442 & 1.437 & 0.001 \\ -0.015 & 0.119 & -0.386 & -0.000 \\ 0.043 & -0.342 & 1.109 & 0.001 \\ -0.009 & 0.074 & -0.241 & -0.000 \\ -0.033 & 0.263 & -0.855 & -0.001 \end{pmatrix} \quad k = 2 : \begin{pmatrix} -0.250 & 0.135 & -1.114 & -0.071 \\ -0.517 & -0.969 & 1.297 & -0.191 \\ 0.150 & 0.270 & -0.346 & 0.055 \\ 0.758 & 0.316 & 1.284 & 0.241 \\ -0.356 & -0.245 & -0.325 & -0.117 \\ 0.216 & 0.492 & -0.795 & 0.083 \end{pmatrix}$$

$$k = 3 : \begin{pmatrix} -0.275 & 0.125 & -1.116 & 0.032 \\ -0.521 & -0.970 & 1.296 & -0.178 \\ 0.293 & 0.328 & -0.333 & -0.540 \\ 0.731 & 0.305 & 1.281 & 0.354 \\ -0.374 & -0.252 & -0.327 & -0.042 \\ 0.146 & 0.464 & -0.801 & 0.373 \end{pmatrix} \quad k = 4 : \overline{X}$$

Il s'agit donc d'une formule de reconstitution, lorsque $k < p$ on obtient une approximation du tableau initial et si $k = p$ on retrouve la matrice initiale X centrée. Pour terminer, on veut ajouter les deux correcteurs initialement écartés de l'étude dans le graphique de l'ACP des individus. On réalise d'abord un centrage de ses deux vecteurs par la moyenne de chaque variable puis on calcule leurs coordonnées dans chaque plan.

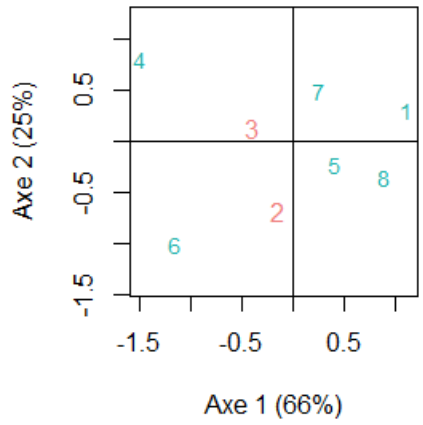


Figure 1.7 – ACP sur les individus (correcteurs) dans le premier plan.

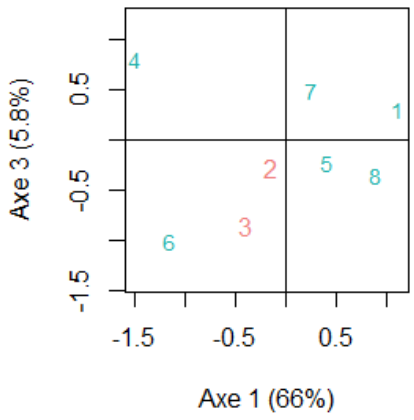


Figure 1.8 – ACP sur les individus (correcteurs) dans le second plan.

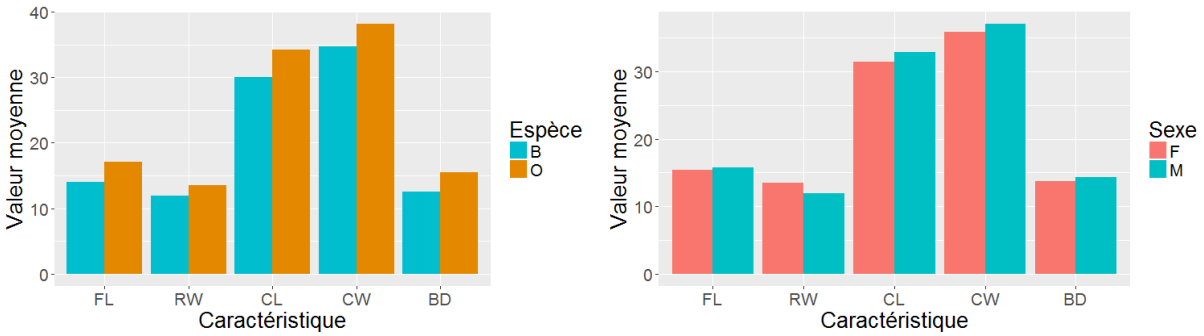
2 Utilisation des outils R

Nous utiliserons ici la fonction `princomp` de l'outil R qui permet d'effectuer une ACP. La fonction `summary` permet d'obtenir les inerties simples et cumulées des différents axes, et `loadings` renvoie les vecteurs propres de l'ACP. Les composantes principales peuvent être retrouvées dans la variable `$scores` de l'objet `princomp`. La fonction `plot` appliquée à un objet `princomp` permet d'obtenir un rendu graphique sur les valeurs des variances des composantes principales, et ainsi de pouvoir les comparer. La fonction `biplot` de la classe `princomp` affiche le plan de représentation des composantes principales. L'option `scale` de `biplot` permet de changer l'échelle et l'option `pc.biplot` mise à la valeur `TRUE` augmente les observations de la racine carré des individus et réduit celle des variables. L'option `choices` est très utile car elle permet de changer les axes de représentation avec par exemple `choices = c(1,3)` qui affiche les axes 1,3.

3 Données crabs

3.1 Statistique descriptive

Le jeu de données contient des informations sur les caractéristiques de 200 crabes. Les crabes sont répartis selon leur sexe (mâle ou femelle) et leur espèce (bleu ou orange). Pour chaque combinaison *sexe/espece* le jeu de données contient 50 individus. Cette répartition équitable permet une analyse statistique plus fiable. Les caractéristiques des crabes sont des longueurs de certaines parties du corps mesurées en millimètres. Elles sont représentées par cinq variables quantitatives : *FL, RW, CL, CW* et *BD*. La dernière donnée des individus qui s'appelle index correspond à un identifiant allant de 1 à 50 pour chaque individu de la sous-catégorie *sexe/espece*. Nous choisissons de représenter les moyennes des caractéristiques des individus sous forme de *barplot* plutôt que sous forme d'un *boxplot*, car même si l'information sur les quartiles est perdue, elle est plus lisible sous cette forme.



(a) Moyennes par espèces. **(b)** Moyennes par sexe.

Figure 3.1 – Moyennes des variables par groupes.

L'espèce d'un individu semble avoir une influence sur ses caractéristiques, les valeurs des variables quantitatives des espèces Bleues sont en moyenne plus petites que les Orange. L'influence du sexe du crabe ne semble, au contraire, pas significative. En effet, les moyennes des mâles sont plus élevées sur quatre des cinq variables mais les valeurs sont trop rapprochées pour en tirer des conclusions.

On peut alors analyser les corrélations entre les différentes variables quantitatives des crabes :

Table 3.1 – Table des corrélations des variables quantitatives des crabes.

	FL	RW	CL	CW	BD
FL	1.00	0.91	0.98	0.96	0.99
RW	0.91	1.00	0.89	0.90	0.89
CL	0.98	0.89	1.00	1.00	0.98
CW	0.96	0.90	1.00	1.00	0.97
BD	0.99	0.89	0.98	0.97	1.00

Toutes les variables sont fortement corrélées. La corrélation est toujours strictement supérieur à 0.88. Les variables quantitatives représentant les longueurs des parties du corps des crabes, il est compréhensible qu'elles soient proportionnelles. Par analogie, il est normal qu'une personne avec de grands bras possède également de grandes jambes. Pour corriger ce phénomène, il est possible de diviser ces valeurs par la variable la plus fortement corrélée, dans notre cas c'est la variable CL et nous la supprimons de l'étude. On obtient ainsi les nouvelles valeurs des corrélations pour les quatre variables restantes.

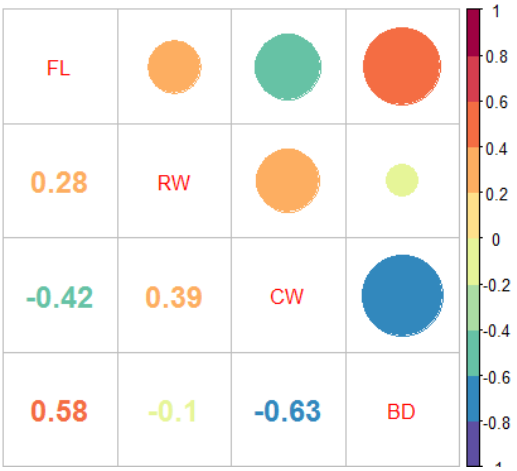
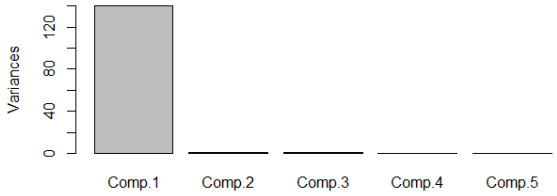


Figure 3.2 – Corrélations des variables après transformation.

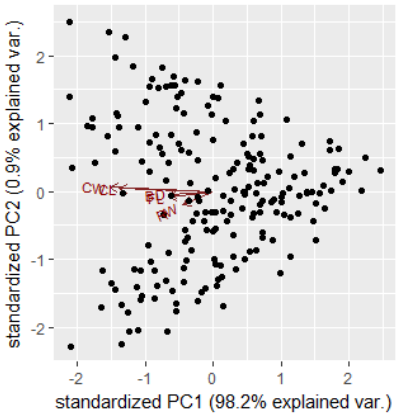
Après transformation, les variables sont nettement moins corrélées. La corrélation la plus forte est maintenant de 0.58. De plus, certaines variables possèdent désormais une corrélation négative entre elles, comme c'est le cas pour les variables *FL* et *CW*.

3.2 Analyse en composantes principales

En effectuant l'ACP sur le jeu de données initial, les graphiques obtenus montrent que toute l'inertie est concentrée sur le premier axe, soit près de 98%. Les échelles totalement déséquilibrées, sur la représentation graphique du nuage de points sur le premier plan factoriel, confirment ce calcul d'inertie. Ce phénomène s'explique par la corrélation très important entre toutes les variables. La solution est d'appliquer au jeu de données l'opération effectuée précédemment, soit de retirer la variable CL en divisant les valeurs des autres variables par celle-ci.



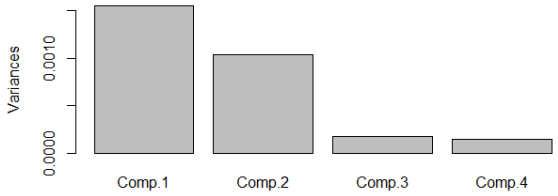
(a) Variances des composantes principales.



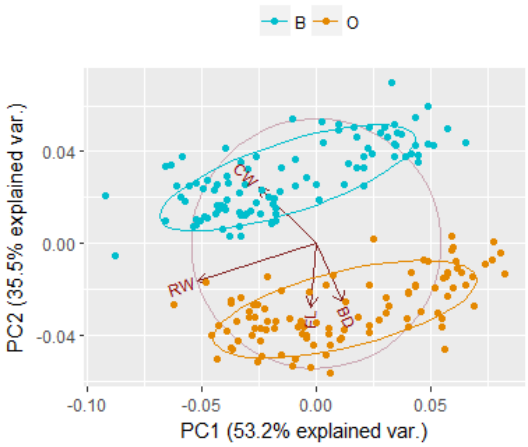
(b) Représentation des individus sur le premier plan factoriel.

Figure 3.3 – Représentation de l'ACP intiale.

L'ACP réalisée sur le jeu de données modifié devient beaucoup plus lisible et permet ensuite de séparer les points selon les deux espèces de crabes, comme cela aurait pu être attendu lors de l'observation initiale des données. On remarque qu'ici, le pourcentage d'inertie de la première composante est beaucoup plus modérée, 53% contre 98% précédemment.



(a) Variances des composantes principales.



(b) Représentation des individus sur le premier plan factoriel.

Figure 3.4 – Représentation de l'ACP après transformation.

4 Données pima

4.1 Statistique descriptive

Le jeu de données *Pima* représente 532 individus diabétiques ou non, décrits par huit variables, dont une seule variable qualitative. La variable Z vaut "1" pour un individu non diabétique et "2" pour un diabétique. Le résumé des valeurs de ces variables selon le facteur « diabète » peut être représenté sous forme de boîte à moustaches :

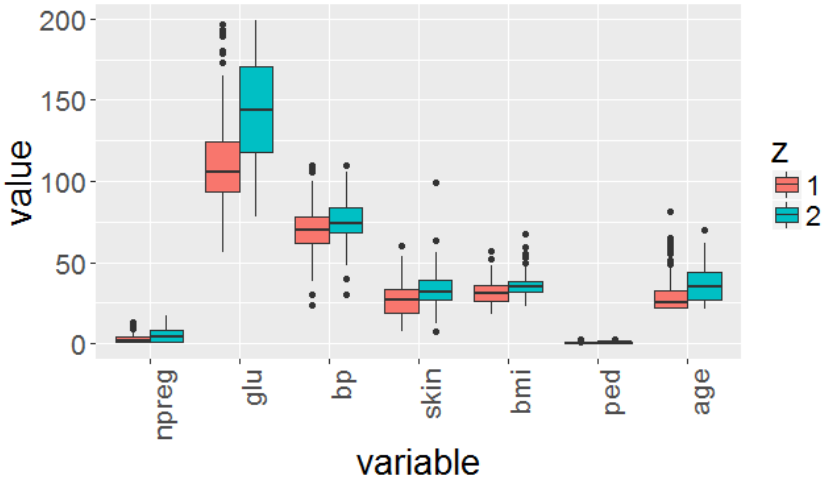
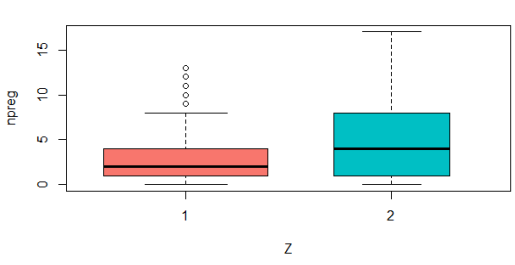
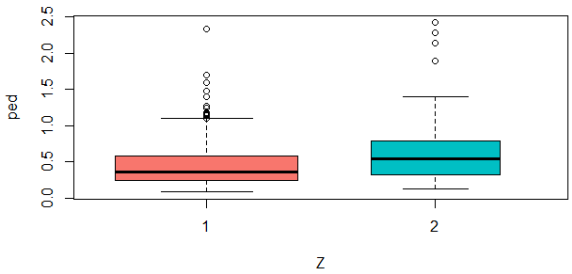


Figure 4.1 – Influence du diabète sur les variables des individus.

Il est possible de faire un zoom sur les variables `npreg` et `ped`, mal représentées sur le graphique précédent dû à l'échelle.



(a) Influence du diabète sur le nombre de grossesses.



(b) Influence du diabète sur la fonction de pedigree du diabète.

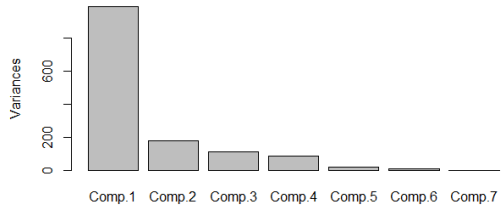
Figure 4.2 – Influence du diabète sur les variables npreg et ped.

Afin de compléter cette représentation des variables, nous décidons de réaliser un test ANOVA sur la variable `skin`, qui représente l'épaisseur du pli cutané au niveau du triceps. En effet, les deux boîtes semblent bien distinctes, mais pas suffisamment pour affirmer graphiquement l'influence significative du diabète sur cette variable. Le test ANOVA nous confirme que l'influence du diabète est significative, on rejette l'hypothèse H_0 au seuil de confiance 1%.

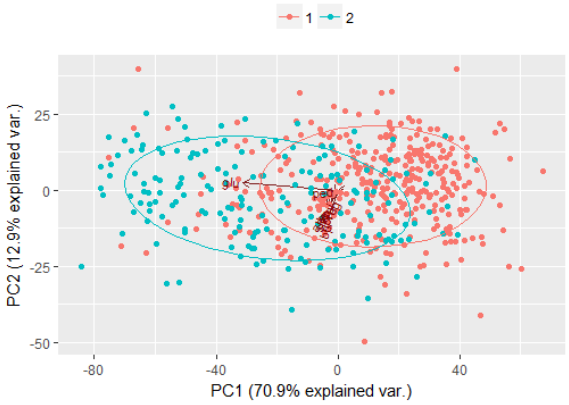
En effet, malgré la ressemblance graphique de ces deux boîtes, on peut expliquer ce rejet par le grand nombre d'individus, qui confirme que cette différence à priori faible est tout de même significative. C'est probablement le cas des autres variables, mais si on observe malgré tout beaucoup de valeurs extrêmes assez éloignées des quartiles $Q1$ et $Q3$.

4.2 Analyse en composantes principales

Nous réalisons une ACP sur le jeu de données *Pima* en séparant les points selon le facteur diabète. La première remarque est qu'une tendance semble se dessiner entre les deux groupes, ce qui tend à les séparer. Cependant, une grande partie des points appartiennent aux deux ellipses qui représentent chacune la zone où se situent la majorité des points du groupe. De plus, certains points sont très distincts des autres, ils correspondent aux valeurs extrêmes révélées précédemment par l'analyse des données. En prenant un point au hasard de cette représentation, il semble possible de déterminer sa probabilité d'appartenance à l'un des deux groupes, mais il serait impossible de l'affirmer avec certitude, à cause des valeurs extrêmes contenus dans le graphe.



(a) Variances des composantes principales.



(b) Représentation des individus sur le premier plan factoriel.

Figure 4.3 – Représentation de l'ACP de Pima.

Conclusion

Lors de ce TP nous avons pu appliquer la statistique descriptive et l'ACP sur trois jeux de données avec des effectifs et des répartitions bien distinctes ce qui a fait ressortir leur aspect complémentaire dans l'extraction d'informations supplémentaires. Le jeu de données *Notes* avait des données très hétérogènes ainsi que des effectifs très différents pour chacun des facteurs et parfois des données manquantes. Les données *Crabs* étaient équitablement réparties selon chaque facteur. Pour terminer, la répartition des données *Pima* selon le facteur diabète était plutôt hétérogène et les effectifs totaux suffisamment importants pour en conclure des analyses statistiques. On retiendra, notamment, que l'application classique de l'ACP ne se suffit pas à elle seule. Il faut parfois effectuer des traitements supplémentaires pour mettre en avant des groupes d'individus.