Statistique descriptive, Analyse en composantes principales

Yanqing ZENG GI05
Minh Tri LÊ GI02

4 avril 2017

Introduction

Le but de ce premier TP est d'une part de décrire puis découvrir des relations au sein de ces données en appliquant des concepts élémentaires de statistique. D'autre part, l'objectif est de réaliser une analyse en composantes principales (ACP) afin d'identifier des similarités ou différences dans les données et de représenter les données dans une dimension réduite sans trop perdre d'information.

1 Statistique descriptive

1.1 Notes

Ce jeu de données comprend les notes de 296 étudiants ayant suivi SY02 au semestre P16. Il comporte 11 variables dont 3 quantitatives et 8 qualitatives.

1.

Table 1 – Résumé des variables du jeu de données *Notes*

	Nature	Domaine de définition	Nombre de valeurs *NA*
Nom	Qualitative	Etu$i$, $i \in [1;296]$ ($\mathbb{N}$)	0
Specialite	Qualitative	{TC,HuTech, ISS, GB, GI,GSU,GM,GSM,GP}	0
Niveau	Qualitative	$[1;6]$ ($\mathbb{N}$)	0
Statut	Qualitative	{UTC, Echange}	0
Dernier diplôme obtenu	Qualitative	{BAC, DUT, AUTRE 1ercycle, CPGE, BTS, DEUG, LICENCE, INGENIEUR, ETRANGER SUPERIEUR, ETRANGER SECONDAIRE, AUTRE DIPLOME SUPERIEUR AUTRE 2E cycle]	6
Note median	Quantitative	$[0;20]$ ($\mathbb{R}$)	3
Correcteur median	Qualitative	Cor$i$, $i \in [1;8]$ ($\mathbb{N}$)	3
Note final	Quantitative	$[0;20]$ ($\mathbb{R}$)	12
Correcteur final	Qualitative	Cor$i$, $i \in [1;8]$ ($\mathbb{N}$)	12
Note totale	Quantitative	$[0;20]$ ($\mathbb{R}$)	12
Resultat	Qualitative ordinale	[A,B,..., F, FX]	12

Le tableau [1]indique une liaison entre les variables *Note median* et *Correcteur median* ainsi que *Note final, Correcteur final, Note totale et Resultat*. On en déduit que si un étudiant n'a pas eu de note au médian (*resp. final*) (pour cause d'absence) alors aucun correcteur du médian (*resp. final*) n'a pu corrigé sa copie. De plus, si un correcteur n'a pas pu corrigé le final alors l'étudiant n'a pas de *note totale* et de *resultat*.

Les 6 valeurs manquantes de *dernier diplome obtenu* correspondent aux 6 étudiants en échange. De plus, on observe que ces 6 étudiants n'ont pas obtenu SY02 ou ont été absents.

Un résumé des statistiques des valeurs quantitatives est en table [4].

2.

Nous recherchons d'abord quelles variables influencent la note totale et par équivalence, la réussite à l'UV.

Les box plot [1]montrent une disparité de la note totale en fonction de la spécialité/niveau et du dernier diplôme obtenu. Au sein d'une même specialité (une même couleur) mais d'un niveau différent, la réussite n'est pas homogène, de même entre les différentes spécialités. Les GSU/TC semblent le mieux réussir contrairement aux GP/GSM/HuTech, cependant il faut noter que les GP, TC et HuTech ne représentent qu'une petite partie des indivus (respectivement 5.4%, 3.7% et 0.3%) donc ce n'est pas très représentatif. Il y a aussi quelques valeurs aberrantes, pour les GI et GSU par exemple.

Nous avons effectué des tests du Chi2 pour la spécialité et le niveau (table [5]). On rejette donc l'hypothèse d'indépendance entre la specialite et le niveau.

Concernant la variable *dernier diplome obtenu*, les CPGE ont le plus de disparités (Étendue = 15.8 soit 79% de 20). Ceux issus de "etranger secondaire" semblent le mieux réussir suivi de BAC et Licence. Némanmoins, il faut savoir que les licences et etranger secondaire ne sont pas quantitativement bien représentés (respectivement 3.1% et 1.7% de la part des individus). Il y a encore quelques valeurs aberrantes pour ceux issus du BAC. Le test du Chi2 (table [5]) ne rejette pas l'hypothèse d'indépendance pour le dernier diplome obtenu.

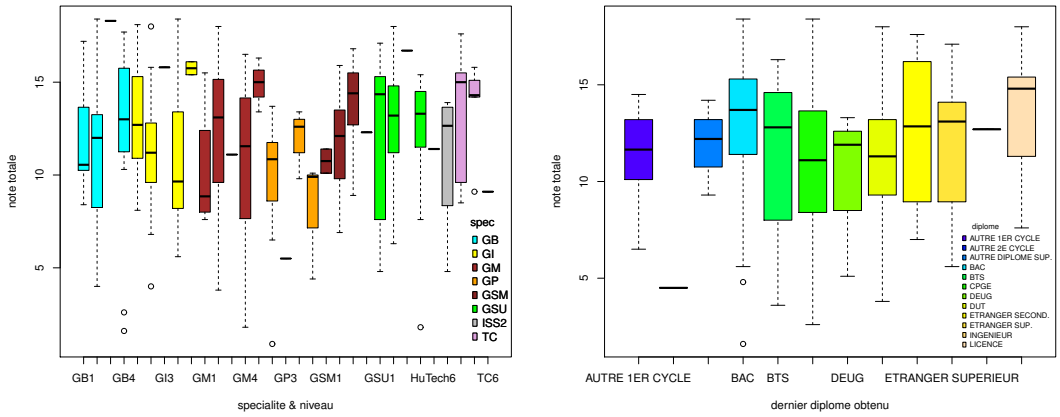

FIGURE 1 – Box plot de la *note totale* en fonction de la *specialite et du niveau* ainsi que du *dernier diplôme obtenu*

Le premier box plot de la figure [2]confirme le fait que les étudiants en échange ne réussissent pas l'UV.

Il y a un peu d'hétérogénéité entre les notes données par les correcteurs median, mais cela est plus marqué pour la note finale ; ce qui se répercute sur la note totale car le dernier box plot ressemble au box plot du correcteur final.

Pour tester si un correcteur influence la note obtenue, nous avons effectué des tests de Student

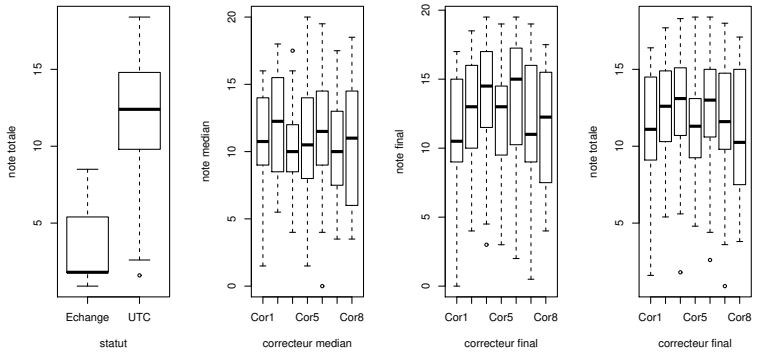

FIGURE 2 – Box plot des notes en fonction du *statut* et du *correcteur*

(table 6). Pour l'examen médian, il ne semble pas y avoir de dépendance entre les correcteurs. En revanche, pour les correcteurs [1,4], [1,6] de l'examen final, l'hypothèse d'indépendance est rejetée : certains correcteurs influencent la note obtenue.

L'intersection des droites moyennes du graphique [3]n'est pas au centre, mais plutôt décalé vers le haut et un peu à droite. Cela indique une meilleure note au final qu'au médian et une réussite générale des étudiants. Il y a une concentration des points dans la partie haute à droite. Ces étudiants ont eu une note au-dessus de la moyenne au médian ou au final ou peut être les deux. À l'inverse, moins d'étudiants ont eu une mauvaise note au médian ou une mauvaise note au final ou peut être les deux.

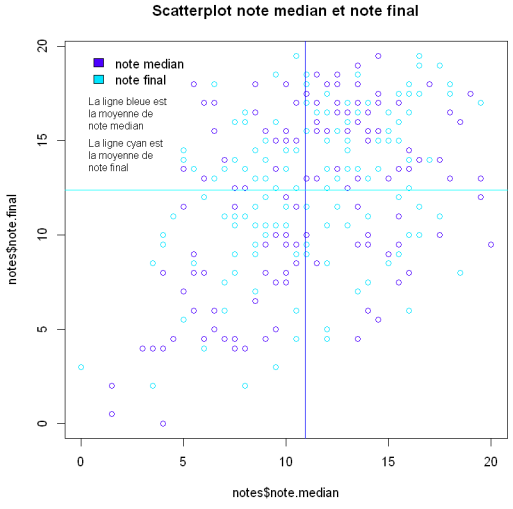

FIGURE 3 – Scatter plot de la *note median* et de la *note final*

Le tableau de corrélation [7]révèle en fait une corrélation plutôt faible (0.386) entre *note median* et *note final* donc il y a peu de lien entre les étudiants ayant eu une bonne note au médian et au

final. La forte corrélation entre *note total* et *note final* (0.912) (resp. *note median* : 0.730) est expliquée par la notation suivante : $note\ total = 0.4 * note\ median + 0.6 * note\ final$.
Ainsi, la réussite ou l'échec au médian n'est pas déterminante de la note finale, mais ces dernières le sont pour la réussite globale.

1.2 Données crabs

Ce jeu de données comprend 200 enregistrements de données sur des crabes. Il comporte 2 variables qualitatives et 5 variables quantitatives. La colonne 'index' est inutile pour cette analyse. Il n'y a aucune valeur *NA* et deux espèces de crabe 'B' (bleu) et 'O' (orange). Un résumé des statistiques des variables quantitatives est en table 8.

1.

Le diagramme en boîte [4]nous montre les relations entre la morphologie et les caractéristiques qualitatives (l'espèce et le sexe). Cela nous donne l'intuition que ces caractéristiques et la morphologie ne sont pas indépendantes. Afin de vérifier la dépendance, nous effectuons un test de Student (Table [9]).

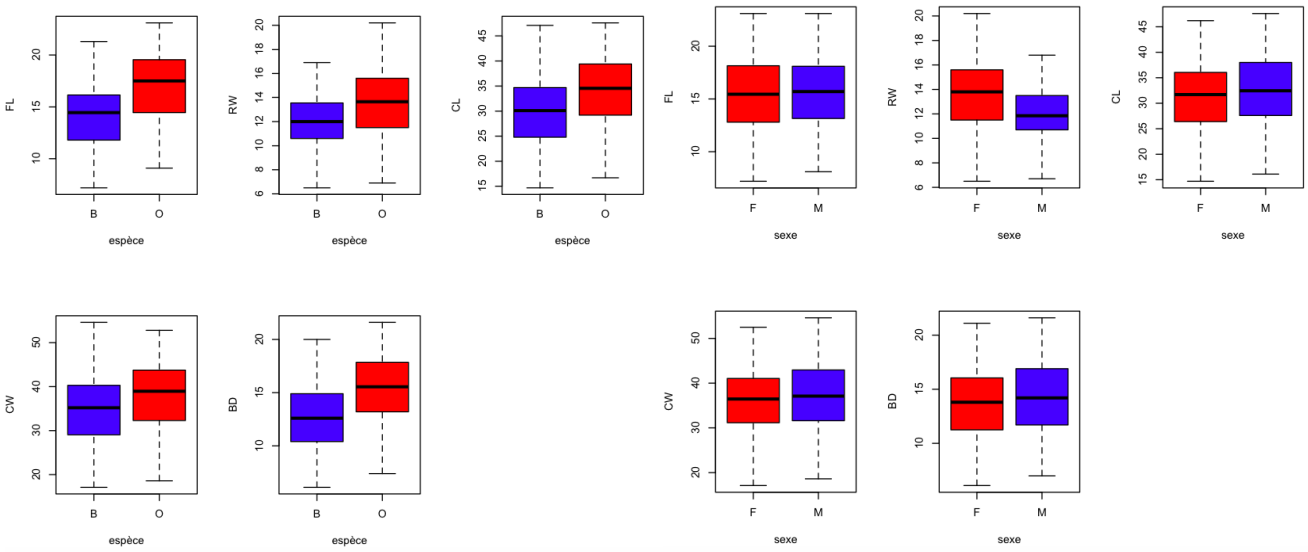

FIGURE 4 – Box plot de la morphologie en fonction de l'espèce (gauche) et du sexe (droite)

Le diagramme [4]nous montre que la distribution des caractéristiques morphologiques de l'espèce 'O' est généralement plus grande que l'espèce 'B', en particulier pour le caractère 'FL' et 'BD'. Cela peut être expliqué par le fait que 'FL' et 'BD' ont les p-values les plus significatives pour le test de Student (Table 9) avec l'espèce.
Les diagrammes en boîte [4]nous montre que la distribution de toutes les variables quantitatives en fonction du sexe est similaire sauf pour le caractère 'RW'. L'individu femelle a à priori un plus grand 'RW' que l'individu mâle. La figure 19 montre que le sexe peut être déterminé en grande partie à partir de 'RW' mais aussi [RW, FL], [RW, CL], [RW,BD] et [RW, CW]. Cela est confirmé par la p-value du test de Student entre le sexe et les variables quantitatives (Table 9) : l'hypothèse d'indépendance est rejetée.

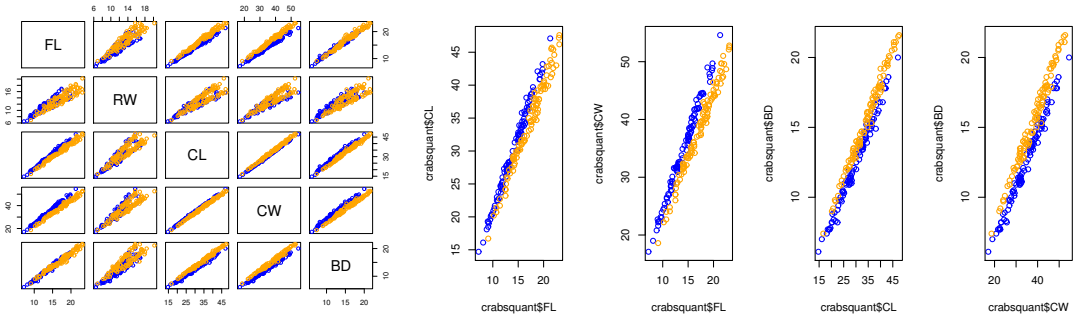

FIGURE 5 – Représentation des individus en fonction de l'espèce ('O' : Orange, 'B' : Bleu) par les variables quantitatives

La figure 5 montre qu'il est possible d'identifier l'espèce en très grande partie avec les variables [FL,CL] et donc [FL,CW] car CL et CW sont très corrélées (voir table 10) De même pour [BD,CW] et donc [BD,CL]. Il y a un recouvrement très faible entre les deux espèces.

2.

La représentation par les paires de variables (Figure 5) nous montre que les 5 variables quantitatives ont une forte corrélation linéaire. Les valeurs de corrélation sont présentées dans la table [10]. La matrice de corrélation montre bien la forte relation linéaire entre les variables, la plus petite valeur est 0.889 ('RW', 'BD'), ce qui est très élevé.
Nous pouvons interpréter cela par le fait que ces variables sont des mensurations de crabes, donc chaque partie du corps d'un crabe est proportionnelle aux autres. 'CL' (Carapace length) et 'CW' (Carapace width) sont corrélées à 0.995 car la longueur et la largeur de la carapace sont proportionnelles.

En conséquence, les individus représentés par ces variables se recouvrent. Il est donc difficile de discerner correctement des groupes en travaillant sur 5 dimensions fortement corrélées.
Pour s'affranchir de ce phénomène, on peut appliquer une analyse en composantes principales (ACP) et réduire les dimensions. On obtient alors des combinaisons linéaires indépendantes des variables, ce qui permet alors de visualiser les individus dans une nouvelle représentation. En observant la figure, il semble qu'une seule dimension peut déjà exprimer une partie de l'information.

1.3 Données pima

1.

Ce jeu de données comprend 532 individus de sexe féminin décrit par huit variables. Il comporte 7 variables quantitatives et 1 variable qualitative.

Un résumé des statistiques des variables quantitatives est en annexe [11]

2.

La figure 6 montre que les individus sont très dispersés entre eux selon les variables quantitatives. La table 12 confirme que les variables sont peu corrélées entre elles sauf pour npreg et age

(0.641), bmi et skin (0.647). Autrement, les valeurs sont inférieures à 0.350. Il y a donc beaucoup de recouvrement sur la représentation des individus diabétiques et non-diabétiques.

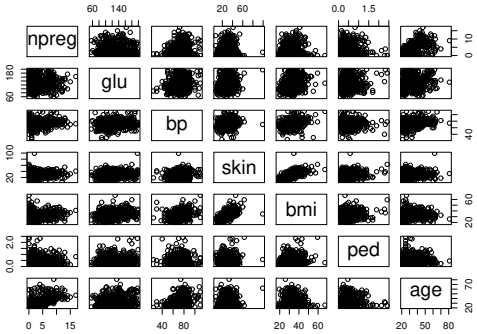

FIGURE 6 – Représentation des individus par les variables quantitatives de 1.3

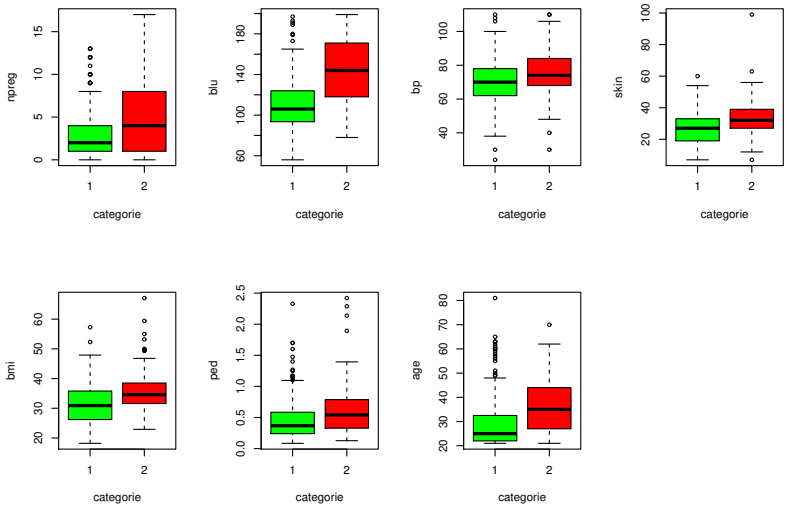

FIGURE 7 – Diagramme en boîte des données pima en fonction du diabète
(Diabétique en rouge, non-diabétique en vert)

La figure 7 nous montre qu'il y a des valeurs aberrantes dans ce jeu de données, en particulier pour l'âge des non-diabétiques. Les individus diabétiques semblent avoir les valeurs les plus élevées par rapport aux non-diabétiques pour tous les indicateurs, surtout pour le glucose.
Pour vérifier l'influence du facteur "diabète", nous avons testé l'hypothèse d'indépendance par le test de Student (table 13)

D'après les résultats du test, on rejette l'hypothèse H_0 : *Les indicateurs sont indépendants de la variable diabète.* Autrement dit, le diabète influence bien les variables quantitatives. La p-value de la variable glu est très significative, cela confirme que le diabète a bien une forte influence sur le taux plasmatique de glucose, ce qui est cohérent (pour le diabète).

Ainsi, il va être difficile de discerner les diabétiques des non-diabétiques en réduisant le nombre de variables avec l'ACP car le diabète influence toutes les variables. De plus, les variables sont en général peu corrélées entre elles donc il n'existe peut-être pas de combinaisons linéaire indépendante avec moins de composantes qu'actuellement.

2 Analyse en composantes principales

2.1 Exercice théorique

1.

Les individus Cor2, Cor3 sont d'abord enlevés car ils ont des valeurs NA.
Après avoir centré les données par colonne, on obtient la matrice X qui prend comme variable : *moy.median, std.median, moy.final, std.final* et les individus *Cor1, Cor4, Cor5, Cor6, Cor7, Cor8* :

$$X = \begin{pmatrix} -0.0058 & -0.1556 & -1.2133 & 0.0679 \\ -0.4795 & -1.0130 & 1.2815 & -0.1723 \\ 0.2654 & 0.3572 & -0.3235 & -0.5436 \\ 0.7858 & 0.2473 & 1.2616 & 0.3617 \\ -0.5917 & -0.0257 & -0.2490 & -0.0705 \\ 0.0258 & 0.5898 & -0.7574 & 0.3568 \end{pmatrix} \tag{1}$$

La matrice de la variance V :

$$V = \frac{1}{n} X^T X = \begin{pmatrix} 0.2114 & 0.1343 & 0.0710 & 0.0455 \\ 0.1344 & 0.2646 & -0.2255 & 0.0453 \\ 0.0710 & -0.2255 & 0.9077 & 0.0127 \\ 0.0455 & 0.0453 & 0.0127 & 0.0988 \end{pmatrix} \tag{2}$$

En diagonalisant la matrice (2), on obtient les valeurs et vecteurs propres de la matrice V :

$$\lambda_1 = 0.97994449, \lambda_2 = 0.36748816, \lambda 3 = 0.08318286, \lambda_4 = 0.05200103$$

La matrice des vecteurs propres U et donc les axes factoriels :

$$U = \begin{pmatrix} -0.0368 & 0.7043 & -0.2332 & 0.6695 \\ 0.2942 & 0.6469 & -0.0943 & -0.6972 \\ -0.9550 & 0.1720 & -0.0206 & -0.2406 \\ -0.0006 & 0.2364 & 0.9676 & 0.0883 \end{pmatrix} \tag{3}$$

Les pourcentages d'inertie expliquée par chacun des axes factoriels : $E_k = \frac{\lambda_1+\lambda_2+...\lambda_k}{\sum_{\alpha=1}^{p} \lambda_\alpha} * 100$:

TABLE 2 – Inertie expliquée par les axes factoriels

	k=1	k=2	k=3	k=4
Inertie expliquée (%)	66.1%	24.8%	5.61%	3.51%
Inertie expliquée cumulée (%) (E_k)	66.1%	90.09%	96.5%	100%

La table 2 montre que les deux premiers axes factoriels contiennent une grande partie de l'inertie expliquée (90%).

2.

La matrice des composantes principales et donc les coordonnées des individus dans le nouvel espace représenté par la figure 8 :

$$C = XU = \begin{pmatrix} 1.1131 & -0.2973 & 0.1068 & 0.4025 \\ -1.5042 & -0.8134 & 0.0142 & 0.0617 \\ 0.4045 & 0.2338 & -0.6149 & -0.0415 \\ -1.1613 & 1.0159 & 0.1174 & 0.0820 \\ 0.2521 & -0.4929 & 0.0773 & -0.3245 \\ 0.8957 & 0.3538 & 0.2993 & -0.1801 \end{pmatrix} \tag{4}$$

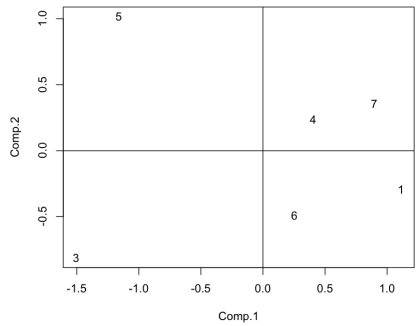

FIGURE 8 – Représentation des six individus dans le premier plan factoriel

3.

La matrice de coordonnées des variables dans le nouvel espace :

TABLE 3 – Coordonnées des variables dans le nouvel espace

	Comp.1	Comp.2	Comp.3	Comp.4
moy.median	-0.079	0.928	-0.146	0.332
std.median	0.566	0.762	-0.053	-0.309
moy.final	-0.992	0.109	-0.006	-0.058
std.final	-0.002	0.456	0.888	0.064

D'après la matrice de coordonnées, on peut déduire que la première composant exprime une majorité de l'information de la variable moy.final, et la deuxième composante exprime bien la variable moy.median et un peu moins std.median. La troisième composante explique uniquement bien la variable std.final.

Plus une variable s'éloigne d'un axe, plus elle est expliquée par celui-ci.

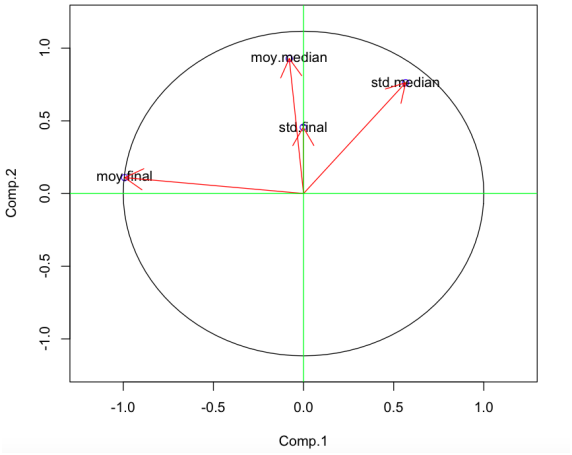

FIGURE 9 – Représentation des 4 variables dans le premier plan factoriel

4.

Quand k=1 :

$$\sum_{\alpha=1}^{1} c_\alpha {u_\alpha}' = \begin{pmatrix} -0.0410 & 0.3275 & -1.0631 & -0.0006 \\ 0.0554 & -0.4425 & 1.4365 & 0.0009 \\ -0.0142 & 0.1190 & -0.3864 & -0.0002 \\ 0.0428 & -0.3416 & 1.1091 & 0.0007 \\ -0.0093 & 0.0742 & -0.2407 & -0.0001 \\ -0.0330 & 0.2635 & -0.8554 & -0.0005 \end{pmatrix}$$

Quand k=2 :

$$\sum_{\alpha=1}^{2} c_\alpha {u_\alpha}' = \begin{pmatrix} -0.2504 & 0.1351 & -1.1142 & -0.0709 \\ -0.5175 & -0.9687 & 1.2967 & -0.1915 \\ 0.1498 & 0.2703 & -0.3461 & 0.0551 \\ 0.7583 & 0.3156 & 1.2838 & 0.2409 \\ -0.3564 & -0.2447 & -0.3255 & -0.1167 \\ 0.2162 & 0.4924 & -0.7946 & 0.0832 \end{pmatrix}$$

Quand k=3 :

$$\sum_{\alpha=1}^{3} c_\alpha {u_\alpha}' = \begin{pmatrix} -0.2753 & 0.1250 & -1.1164 & 0.0324 \\ -0.5208 & -0.9700 & 1.2964 & -0.1778 \\ 0.2932 & 0.3282 & -0.3335 & -0.5400 \\ 0.7309 & 0.3045 & 1.2814 & 0.3545 \\ -0.3745 & -0.2520 & -0.3271 & -0.0418 \\ 0.1464 & 0.4642 & -0.8008 & 0.3728 \end{pmatrix}$$

Quand k=4 :

$$\sum_{\alpha=1}^{4} c_\alpha {u_\alpha}' = \begin{pmatrix} -0.0058 & -0.1556 & -1.2136 & 0.0679 \\ -0.4795 & -1.0130 & 1.2815 & -0.1723 \\ 0.2654 & 0.3572 & -0.3235 & -0.5436 \\ 0.7858 & 0.2473 & 1.2616 & 0.3617 \\ -0.5917 & -0.0257 & -0.2490 & -0.0705 \\ 0.0258 & 0.5898 & -0.7574 & 0.3568 \end{pmatrix}$$

On trouve que c'est la matrice des données originales : cela permet de reconstituer X. Autrement dit, $CU^T = X$, cela vérifie bien la conclusion de la question 1 : $E_4 = 100\%$

5.

Le code en annexe [20] permet de remplacer les valeurs manquantes par la moyenne des variables correspondantes.

En réitérant la procédure, on obtient les valeurs propres [14], le premier et second plan factoriel [10]. On remarque qu'il faut prendre en compte les trois premiers axes pour arriver à plus de 90% contrairement à avant où deux axes suffisaient. Ainsi, l'importance de l'axe 1 est un peu diminuée au profit des axes 2 et 3.

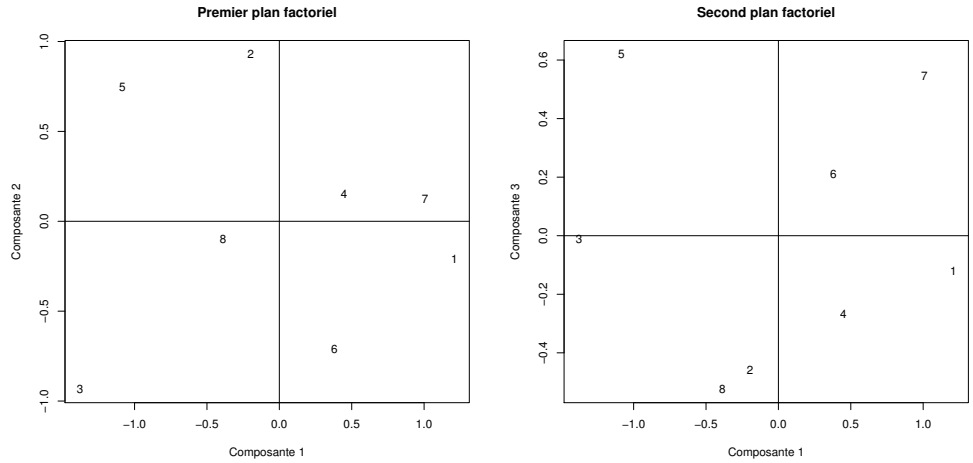

FIGURE 10 – Premier et second plan factoriel

Les individus du premier plan factoriel 10 sont situés d'une manière similaire au premier plan factoriel de la figure 8 sans les individus 8 et 2. Ici, l'axe 2 et 3 explique bien l'individu 2 alors que seul l'axe 3 explique bien l'individu 8.

2.2 Utilisation des outils R

1. Retrouver tous les résultats avec princomp

On obtient l'ACP avec la fonction : **corr.pr<-princomp(data)** qui retourne un objet et affiche l'écart type pour chaque composantes.

On obtient l'écart type des composantes principales (donc la racine des valeurs propres), la proportion de la variance et la proportion cumulée par : **summary(corr.pr)**

Les vecteurs propres ou axes factoriels sont accessibles par **corr.pr$loadings**.

On obtient les coordonnées des individus par les composantes principales dans la variable **corr.pr$scores**.

plot(corr.pr) affiche un histogramme des variances de chaque composante.

biplot(corr.acp) affiche sur le même graphique les individus et les variables sur le premier plan factoriel.

biplot.princomp est la fonction générale de **biplot**. Elle peut prendre plusieurs arguments :

— *choices* : permet de choisir les axes principaux à représenter

— *scale* : permet de mettre les données à l'échelle (entre 0 et 1)

Nous avons retrouvé les résultats obtenus en 2.1 avec ces fonctions.

2. Les resultats de princomp

Avec la fonction plot, on obtient les variance exprimées par les composantes principales. Avec la fonction biplot, on obtient la représentation des individus et des variables qui combine les représentations trouvées en [8] et [9] (Les axes sont orientés dans le sens opposé).

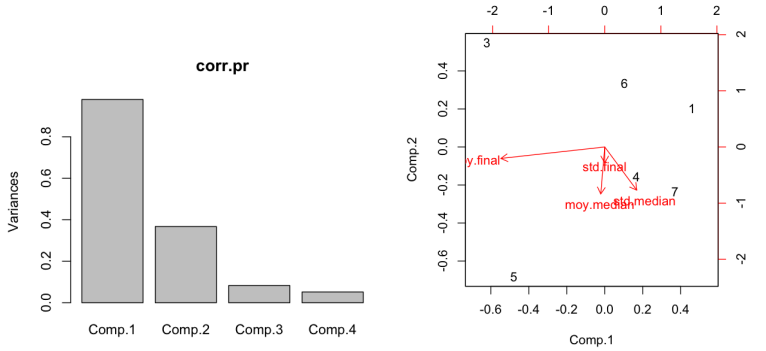

FIGURE 11 – Le résultat des fonctions plot et biplot

2.3 Données Crabs

Sans traitement préalable, on effectue un ACP sur les données. Les variances exprimées et la représentation du biplot sont présentées par les figures suivantes [12] :

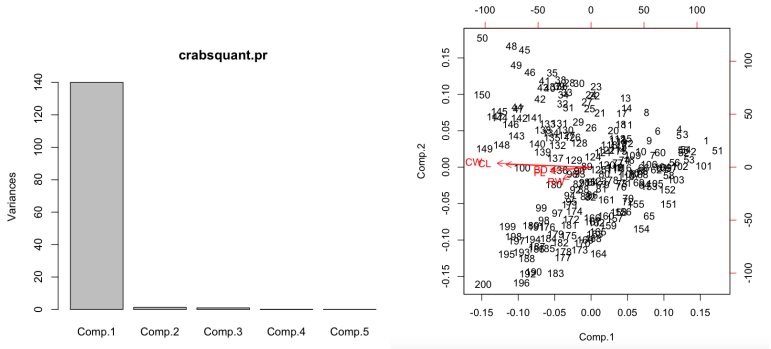

FIGURE 12 – Variances exprimées et la représentation des individus

On observe que la première composante contient une très grande partie de l'information, car elle a une inertie expliquée de 98.2%. (table 15), alors que la deuxième composante explique moins d'1% de la variance totale.
Cela signifie que les individus varient très fortement selon une composante uniquement. Comme nous l'avons expliqué en 1.2, les mensurations des crabes sont toutes corrélées car elles sont proportionnelles entre elles. La première composante est un indicateur de la taille et donc de la forme des crabes. De plus, la très faible variance de la deuxième composante confirme que les crabes restent proportionnels à leurs mensurations pour toutes les espèces et sexes car les individus sont uniquement expliqués par la première composante.

Toutes les variables du biplot 12 sont expliquées par la première composante mais très peu par la deuxième sauf pour RW qui est en plus légèrement expliquée par la deuxième composante.

Cela est expliqué par le fait que les variables sont fortement corrélées comme nous l'avons vu en 1.2.
En particulier, on observe que les variables [CL,CW] ; [FL,BD] sont regroupées ensemble et que RW est proche de [FL,BD] mais quelque peu distante. Autrement dit, chaque groupe de variables exprime un attrait différent sur les individus. Cela explique que l'on ait pu identifier une bonne partie des espèces ou le sexe avec une combinaison de [CL ou CW] avec [FL ou BD] ou avec RW car ces caractéristiques ne sont pas regroupées ensemble.

De plus, il n'est pas possible d'identifier visuellement des groupes d'individus sur le biplot.

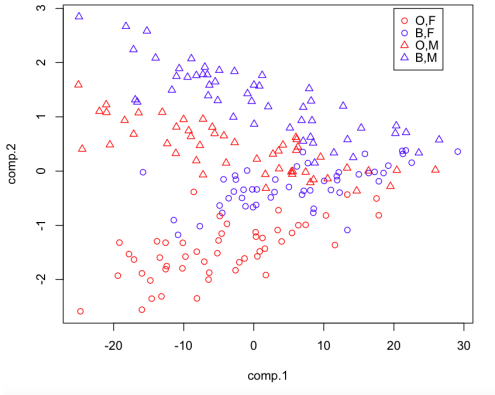

FIGURE 13 – Visualisation dans le premier plan factoriel

Avec de grandes variances exprimées par la première et la deuxième composante principale, on pouvait identifier des catégories, mais ce n'est pas le cas dans ce jeu de données. Nous n'avons pas une claire frontière entre l'espèce 'O' et 'B'.

La figure nous montre que le sexe empêche la distinction des groupes. Dans le tableau de p-value (table 9), on trouve une dépendance entre RW et sexe. Une solution est donc d'enlever la variable RW.

D'après la figure 13, on peut identifier le sexe sachant l'espèce de l'individu selon la deuxième composante principale. Il y a une certaine frontière entre les triangles et les cercles.

Aussi, comme la forme d'un crabe est proportionnelle à toutes ses mensurations chaque combinaison d'espèce et de sexe, une autre solution serait de diviser chaque mensuration par la taille totale de ses membres (donc la somme de ses mensurations). Nous effectuons ainsi une mise à l'échelle des individus.

2.

Après avoir enlevé la variable RW, on reconduit une ACP. La visualisation est présentée par le diagramme suivant :

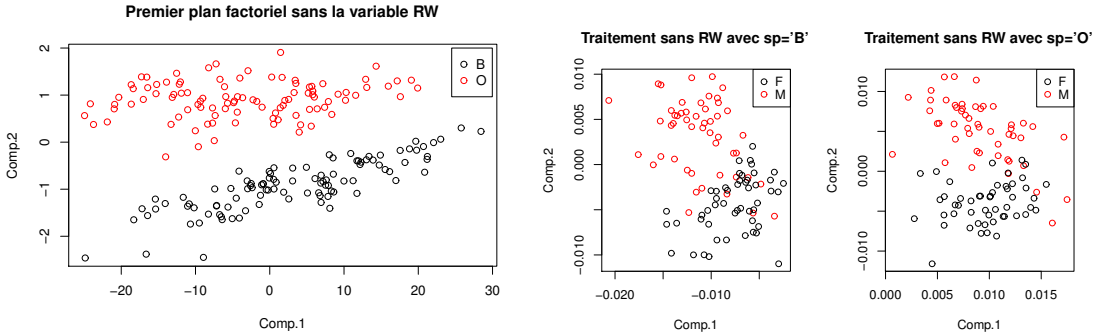

FIGURE 14 – Représentation des individus après avoir enlevé RW

La figure 14 montre qu'il y a une frontière claire entre l'espèce 'B' et 'O' qui permet donc d'identifier ces individus. De plus, en connaissant l'espèce de l'individu, on a une frontière entre les différents sexes.

Avec la solution de mise à l'échelle, on obtient l'inertie expliquée suivante :

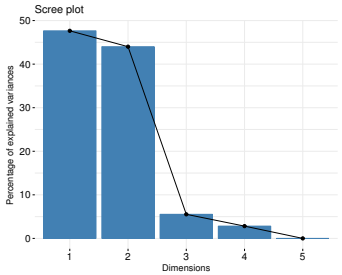

FIGURE 15 – Inertie expliquée de l'ACP

Dans ce cas, la variance de la première composante est bien distribuée à la deuxième et un peu à la troisième composante comparativement à [12].

Les individus sont représentés par les biplots suivants : (species à gauche, sexe à droite).

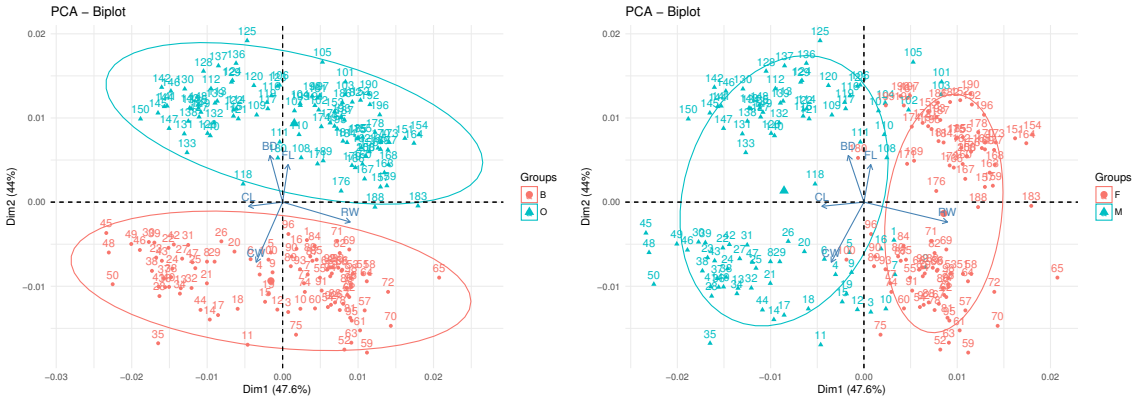

FIGURE 16 – Représentation des individus et variables dans le nouvel espace

D'après la figure 16, nous en déduisons que l'espèce est facilement discernable selon la composante 1 et 2. L'ellipse du biplot pour l'espèce est fixée de sorte à contenir les individus à taux de confiance de 95%.

De même, la figure 16 montre que le sexe est facilement identifiable par la composante 1 et 2 avec un recouvrement tout de même plus visible que pour le biplot du sexe. L'ellipse est paramétrée à un taux de 68%. En fait, en combinant ces deux représentations, nous pouvons maintenant distinguer 4 groupes correspondant aux combinaisons de species et de sex.

On remarque que les variables représentées ne sont plus alignées dans la même direction : RW et CW sont presque orthogonaux.

2.4 Données Pima

Nous avons effectué l'ACP sans traitement préalable et il semble difficile de distinguer visuellement les groupes dans le premier plan factoriel [17]. Par contre, on peut voir une frontière un peu plus claire sur la représentation des trois premiers axes principaux [18].

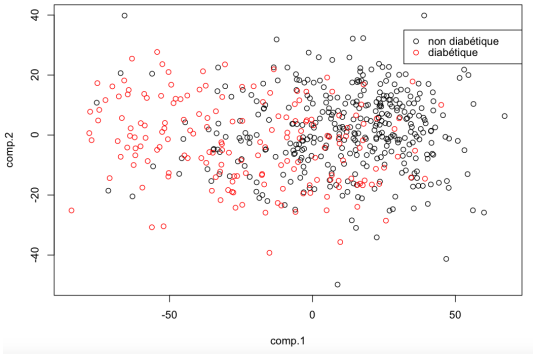

FIGURE 17 – Représentation des individus dans le premier plan factoriel
(Diabétique : rouge, Non diabétique : noir)

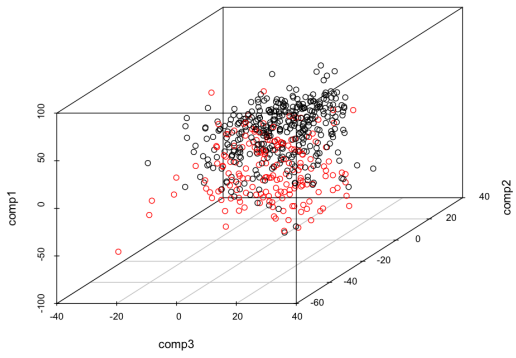

FIGURE 18 – Représentation de pima dans 3 dimensions
(Diabétique : rouge, Non diabétique : noir)

Les corrélations des variables dans la question 1.3 montrent que les variables quantitatives sont assez indépendantes les unes des autres.

Les deux premières composantes principales (table 16) n'expriment pas assez d'information. Par conséquent, on a besoin de plus de dimensions pour pouvoir discerner les groupes d'individus.

Ainsi, comme nous l'avions pressenti en 1.3, il est dans ce cas difficile pour l'ACP de trouver des composantes principales réduisant assez le nombre de dimensions jusqu'à moins de 3 pour distinguer les deux catégories de patientes.

Conclusion

Ce TP nous a permis d'appliquer des concepts de base de statistiques afin d'avoir une première vue sur divers jeux de donnée tant au niveau du nombre de variables qu'au degré de corrélation entre-elles. Nous avons ensuite appliqué une ACP afin d'approfondir notre analyse sur ces jeux de données. Cela nous a permis d'identifier visuellement des groupes et constitue une première approche à l'apprentissage non supervisé.

Références

[1] Principal component analysis in R : prcomp() vs. princomp() - R software and data mining

Annexes

Table 4 – Descriptif statistique des variables quantitatives de 1.1

	Min	1st Qu.	Median	Moyenne	3rd Qu.	Max	Std. dev.
Note median	0	8	11	10.92	14	20	3.82
Note final	0	9.5	13	12.38	16	19.5	4.25
Note total	0.9	9.78	12.3	11.85	14.80	18.40	3.44

Table 5 – Tableau des p-value du test du Chi2 entre les variables qualitatives de 1.1 et le *resultat* pour l'hypothèse d'indépendance

	resultat	
specialite	2.08e-2	Rejetée
niveau	3.52e4-	Rejetée
dernier diplome obtenu	2.57e-1	Non rejetée
correcteur median	6.96e-1	Non rejetée
correcteur final	7.31e-1	Non rejetée

Table 6 – Tableau des p-value du test de Student pour l'hypothèse d'indépendance entre le correcteur et l'examen corrigé (Examen médian à gauche, final à droite)

	Cor1	Cor2	Cor4	Cor5	Cor6	Cor7	Cor8	Couple avec hypothèse rejetée
Cor1	1							\
Cor2	0.182	1						\
Cor4	0.605	0.0117	1					\
Cor5	0.790	0.216	0.333	1				\
Cor6	0.435	0.533	0.096	0.556	1			\
Cor7	0.554	0.018	0.878	0.312	0.105	1		\
Cor8	0.979	0.243	0.625	0.832	0.499	0.575	1	\

	Cor1	Cor3	Cor4	Cor5	Cor6	Cor7	Cor8	Couple avec hypothèse rejetée
Cor1	1							\
Cor3	0.130	1						\
Cor4	0.030	0.301	1					(Cor4, Cor1)
Cor5	0.416	0.345	0.066	1				\
Cor6	0.038	0.344	0.984	0.088	1			(Cor6, Cor1)
Cor7	0.395	0.426	0.097	0.932	0.120	1		\
Cor8	0.738	0.302	0.092	0.708	0.105	0.671	1	\

TABLE 7 – Tableau de corrélation des variables quantitatives de 1.1

	Note median	Note final	Note totale
Note median	1		
Note final	0.386	1	
Note total	0.730	0.912	1

TABLE 8 – Descriptif statistique des variables quantitatives de 1.2

	Min	1st Qu.	Median	Moyenne	3rd Qu.	Max	Std. dev.
FL	7.20	12.90	15.55	15.58	18.05	23.10	3.50
RW	6.5	11.00	12.80	12.74	14.30	20.20	2.57
CL	14.70	27.27	32.10	32.11	37.23	47.60	7.12
CW	17.1	31.50	36.80	36.41	42.00	54.60	7.87
BD	6.1	11.40	13.90	14.03	16.60	21.60	3.42

TABLE 9 – Tableau des p-value du test de Student entre les variables quantitatives et qualitatives de 1.2

	Espèce		Sexe	
FL	8.970e-11	Dépendant	5.426e-01	Indépendant
RW	5.306e-06	Dépendant	2.862e-05	Dépendant
CL	3.468e-05	Dépendant	1.390e-01	Indépendant
CW	2.109e-03	Dépendant	2.949e-01	Indépendant
BD	4.065e-10	Dépendant	2.064e-01	Indépendant

TABLE 10 – Corrélation entre les caractériques morphologiques

	FL	RW	CL	CW	BD
FL	1				
RW	0.907	1			
CL	0.979	0.893	1		
CW	0.965	0.900	0.995	1	
BD	0.988	0.889	0.983	0.968	1

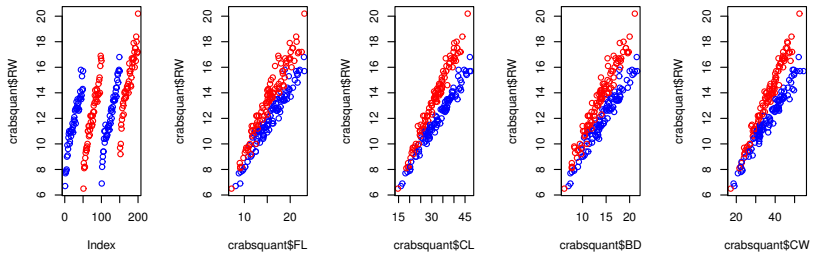

FIGURE 19 – Représentation des individus en fonction du sexe ('F' : rouge, 'M' : Bleu) par l'index et les variables quantitatives identifiants fortement l'espèce

TABLE 11 – Descriptif statistique des variables quantitatives de 1.3

	Min	1st Qu.	Median	Moyenne	3rd Qu.	Max	Std. dev.
npreg	0	1	2	3.52	5	17	3.31
glu	56	98.75	115	121.03	141.25	199	31
bp	24	64	72	71.51	80	110	12.31
skin	7	22	29	29.18	36	99	10.52
bmi	18.20	27.88	32.80	32.89	36.90	67.10	6.88
ped	0.09	0.26	0.42	0.50	0.66	2.42	0.34
age	21	23	28	31.61	38	81	10.76

TABLE 12 – Corrélation entre les variables quantitatives

	npreg	glu	bp	skin	bmi	ped	age
npreg	1						
glu	0.125	1					
bp	0.205	0.219	1				
skin	0.095	0.227	0.226	1			
bmi	0.009	0.247	0.307	0.647	1		
ped	0.007	0.166	0.008	0.119	0.151	1	
age	0.641	0.279	0.347	0.161	0.073	0.072	1

TABLE 13 – Tableau des p-value du test de Student en fonction du facteur diabète

	Résultat du test de Student	
npreg	1.617e-07	dépendant
glu	2.362e-28	dépendant
bp	3.103e-05	dépendant
skin	4.862e-09	dépendant
bmi	2.887e-12	dépendant
ped	9.266e-07	dépendant
age	1.055e-12	dépendant

```
for(i in 1:ncol(corr.acp)){
  corr.acp[is.na(corr.acp[,i]), i] <- mean(corr.acp[,i], na.rm = TRUE)
}
```

FIGURE 20 – Code R pour remplacer des valeurs manquantes par la moyenne des colonnes

TABLE 14 – Valeurs propres et inertie expliquée par les axes factoriels [2.1]

	λ_1	λ_2	λ_3	λ_4
Valeur	0.761	0.362	0.162	0.120
Inertie expliquée (%)	54.2	25.8	11.5	8.50
Inertie expliquée cumulée (%) (E_k)	54.2	79.9	91.5	100.0

TABLE 15 – Valeurs propres et inertie expliquée par les axes factoriels

	λ_1	λ_2	λ_3	λ_4	λ_5
Valeur	140.00	1.29	1.00	0.13	0.078
Inertie expliquée (%)	98.2	0.91	0.70	0.09	0.05
Inertie expliquée cumulée (%) (E_k)	98.2	99.2	99.95112	100.0	100.0

TABLE 16 – Inertie expliquée par les composantes principales

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Standard deviation	31.4579	13.4357	10.6186	9.2122	4.5690	2.4770	3.358e-01
Proportion of Variance	0.7094	0.1294	0.0808	0.0608	0.0149	0.0043	8.086e-05
Cumulative Proportion	0.7095	0.8389	0.9197	0.9806	0.9955	0.99991	1.0000