Manon Beuque
Tanguy Le Floch
Valentin Montupet

TP1 - Statistique descriptive, Analyse en composantes principales

UV : **SY09**

Semestre : **P17**

Table des matières

1. Statistique descriptive

L'objectif de ce TP de SY09 est de pratiquer deux grandes notions. Il vise en premier lieu à permettre aux étudiants de mettre en place des éléments de statistique descriptive en s'appropriant les commandes de R (re)découvertes lors de la première séance. Puis d'effectuer à nouveau les analyses précédentes en utilisant un outil particulier : l'analyse en composantes principales. Cette démarche s'effectuera sur trois sets de données que sont les notes de l'UV SY02 en P2016, Crabs et Pima.

1.1 Notes d'UV

Ce premier jeu de données sous forme d'un fichier csv contient des informations au sujet d'étudiants ayant suivi l'UV SY02 dispensée à l'UTC durant le semestre de Printemps 2016.

Ces données représentent 296 individus (étudiants) pour lesquels on dispose des valeurs de 11 variables :

- le **nom** (anonymisé), variable qualitative nominale,
- sa **spécialité**, variable qualitative nominale ainsi que son **niveau**, variable qualitative ordinale,
- le **statut** de l'étudiant (« En échange » ou « étudiant UTC ») variable qualitative nominale,
- le **dernier diplôme obtenu**, variable qualitative nominale
- les **notes du médian, du final** ainsi que la **note globale** de l'UV qui sont des variable quantitatives discrètes,
- les **correcteurs anonymisés du médian et du final**, variables qualitatives nominales
- le **résultat** obtenu par l'étudiant à l'UV, variable qualitative ordinale.

Analyse des anomalies :

- Dernier Diplôme Obtenu N.A => Pas d'information pour les étudiants étrangers
- Correcteur et notes absents => L'étudiant n'as pas passé l'examen
- Correcteurs manquants d'un examen => Le correcteur n'a corrigé aucune copie pour cet examen (Corr3 pour le médian, Corr2 pour le final)

Une première manière d'aborder ces données serait d'en observer le sommaire (Annexe A.1). On peut notamment remarquer que pour les variables qualitatives, certaines valeurs ont très peu d'individus (HuTech, TC par exemple). Il est donc nécessaire de porter attention à ne pas biaiser l'analyse en utilisant des résultats uniquement en proportion.

Après avoir décrit les différentes variables, nous pouvons désormais émettre des suppositions sur celles étant *a priori* liées.

1.1.1 L'influence des notes du médian sur les notes du final

Nous faisons un test de corrélation linéaire entre les deux colonnes correspondant aux notes du médian et du final : on obtient 0.43 ce qui traduit l'absence de corrélation entre les deux examens.

1.1.2 L'influence du dernier diplôme obtenu sur le résultat

Pour étudier l'influence du dernier diplôme obtenu sur le résultat, nous commençons par remplacer les notes par des booléens (TRUE or FALSE) pour signifier l'attribution ou non de l'UV. Ensuite on créée une table de contingence en fonction du dernier diplôme obtenu :

	notesR	
	FALSE	TRUE
AUTRE 1ER CYCLE	1	5
AUTRE 2E CYCLE	1	0
AUTRE DIPLOME SUPERIEUR	1	2
BAC	15	91
BTS	3	4
CPGE	16	23
DEUG	1	2
DUT	32	57
ETRANGER SECONDAIRE	1	3
ETRANGER SUPERIEUR	5	6
INGENIEUR	0	1
LICENCE	2	7

	notesR	
	FALSE	TRUE
BAC	15	91
CPGE	16	23
DUT	32	57
ETRANGER SUPERIEUR	5	6

FIGURE 1.1 – A gauche : Tableau de Contingence de la réussite à l'UV en fonction du dernier diplôme obtenu. A droite : Tableau de Contingence exempté des éléments à trop faible effectif

On remarque figure 1.1 que certaines valeurs sont inférieures à 5, ce qui rendrait un test du χ^2 peu valide, nous choisissons de n'en considérer que certaines. On réalise ensuite le test pour obtenir une p-value de 0.0004587, ce qui indique une forte corrélation entre dernier diplôme obtenu et la réussite à l'UV !

1.1.3 L'influence du correcteur sur la note

Dans cette partie, nous souhaitons étudier s'il existe une corrélation entre le correcteur et la note obtenue. Pour ce faire, nous allons mener notre étude sur le médian et le final séparément.

Avant de commencer, deux remarques :

- Lors du médian, le correcteur 3 n'était pas présent et n'a noté aucune copie. De la même manière, le correcteur 2 n'a pas participé à la notation de l'examen final.
- Pour chaque élève, les correcteurs du médian et du final sont différents SAUF pour le correcteur 1 qui a évalué les mêmes élèves aux deux examens.

En effectuant l'affichage des diagrammes en boîtes des notes du médian et des notes du final par correcteur, on s'aperçoit que les intervalles de notes sembles proches (Annexe A.2). Afin de déterminer mathématiquement l'indépendance ou non de deux variables qualitatives, nous effectuons un test du χ^2 de contingence. Rappelons les conditions à réunir pour que notre test soit fiable :

Un test du χ^2 s'applique uniquement sur des tableaux de contingence :

- ayant au moins 2 lignes et 2 colonnes,
- contenant des valeurs positives entières,
- ayant au moins 60 observations au total,
- ayant au minimum 5 observations par cases du tableau et/ou dans le tableau des effectifs théoriques.

Si nous créons directement le tableau de contingence la dernière condition n'est pas respectée. Il est donc nécessaire de regrouper les notes en intervalles afin d'obtenir au moins 5 observations pour chaque classe.

	Cor1	Cor2	Cor4	Cor5	Cor6	Cor7	Cor8
[0,10[	8	16	22	20	13	24	11
[10,15[	12	15	25	19	24	17	9
[15,20]	4	17	2	10	12	8	5

FIGURE 1.2 – Tableau de contingence de l'examen médian après transformation

Comme nous pouvons l'observer sur la figure 1.2, il existe des cases du tableau ayant

moins de 5 observations. Bien qu'ayant construit un intervalle $[0, 10[$ plus grand que les deux autres, il aurait fallu rassembler aussi les intervalles $[10, 15[$ et $[15, 20[$ en un unique intervalle, mais nous aurions subi une perte d'information trop importante. Il faudra donc nuancer le résultat du test sur ces données. Nous pouvons désormais effectuer nos tests sur nos deux tableaux de contingence.

Suite à deux test du χ^2, on obtient pour l'examen médian et l'examen final des p-values respectives de 0.0417 et 0.126. On rappelle qu'en général, on rejette l'hypothèse d'indépendance des 2 variables si la p-value est inférieure à 0.05. Dans notre cas, il semblerait que la note du médian soit influencée par le correcteur. Cette information est à prendre avec précaution car l'une des conditions n'était pas totalement remplie. Pour le final, il semble clairement qu'il y ait indépendance entre les variables.

Effectuons un dernier test sur l'ensemble des notes. Pour ce faire, ajoutons le correcteur 3 au tableau du médian, le correcteur 2 au tableau du final, et additionnons les deux tableaux de contingence et effectuons un test du χ^2 dessus. Nous obtenons une p-value de 0.17 donc nous acceptons l'hypothèse d'indépendance entre les correcteurs et les notes.

1.2 Données crabs

Dans cette partie, nous analysons un jeu de données sur deux espèces de crabes réparties en 100 individus chacune. Les 200 individus sont représentés par 8 variables : 3 qualitatives (« sp » pour espèce, « sex » pour le sexe (M ou F) et un index) et 5 variables quantitatives.

Pour répondre à l'interrogation des « différences de caractéristiques morphologiques, en particulier selon l'espèce ou le sexe », nous nous proposons de tracer les différentes variables et de discriminer les individus en deux sur les graphiques. Pour les espèces, on représentera avec une couleur les individus d'espèce B et d'une autre ceux de l'espèce O. On fera de même pour le sexe.

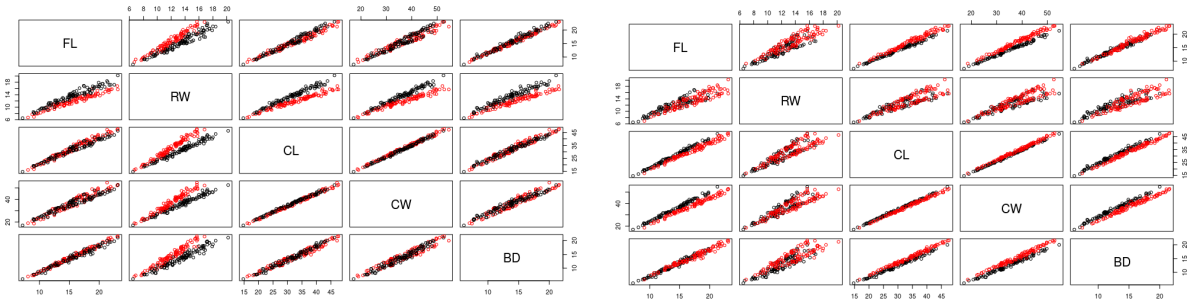


FIGURE 1.3 – Représentation des variables du dataset crabs. A gauche : différentiation selon le sexe. A droite : différentiation selon l'espèce

On voit très clairement figure 1.3 qu'il n'est pas possible de différencier le sexe ou l'espèce d'un individu selon les autres variables disponibles. Cependant, on remarque graphiquement une forte corrélation pour l'ensemble des variables quantitatives. Ce résultat est logique car elles représentent toutes des mesures du corps des différents crabes, qui sont proportionnelles entre elles, sinon les animaux seraient difformes.

	FL	RW	CL	CW	BD
FL	1.000				
RW	0.907	1.000			
CL	0.979	0.893	1.000		
CW	0.965	0.900	0.995	1.000	
BD	0.988	0.889	0.983	0.968	1.000

TABLE 1.1 – Tableau récapitulatif des corrélations entre les 5 variables quantitatives, confirmant l'observation graphique faite précedemment

Quel traitement est-il possible d'appliquer aux données pour s'affranchir de ce phénomène de corrélation ?

En calculant les moyennes des corrélations pour chaque variable (en excluant les cases du type « FL*FL » qui donne évidemment une corrélation de 1), on observe que c'est la variable CL qui est la plus corrélée aux autres avec un coefficient moyen de 0.9624528.

On peut donc diviser l'ensemble des variables par CL afin de décorréler les variables entre elles. Dès lors, nous pouvons visualiser de nouveaux les graphiques précédents.

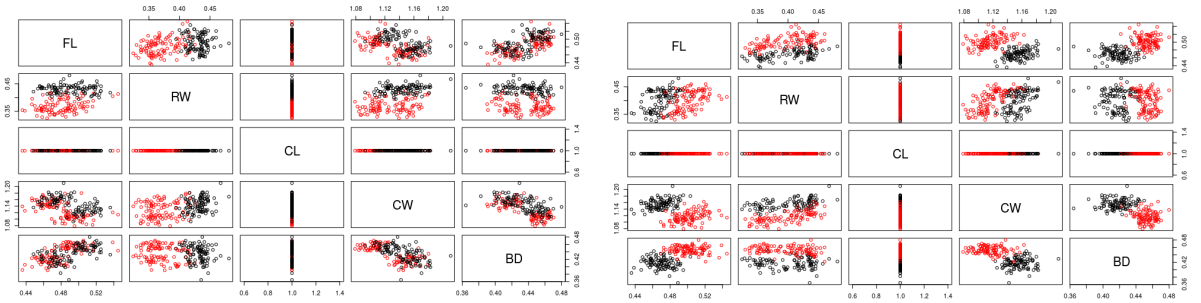


FIGURE 1.4 – Représentation des variables décorrélées du dataset crabs. A gauche : différentiation selon le sexe. A droite : différentiation selon l'espèce

Pour la différenciation selon l'espèce, on distingue très clairement des clusters pour certains rapports entre variables, que l'on pourrait séparer par des droites linéaires (à l'aide d'un procédé de classification, par exemple un SVM), comme le rapport CW/BD ou encore le rapport FL/BD. On peut donc distinguer très clairement les différences entre espèces en assimilant un rassemblement de points à une même espèce.
Concernant la discrimination par sexe, le résultat est plus nuancé : on peut observer des clusters se chevauchant légèrement pour les rapports RW/BD, RW/FL et RW/CW. Pour d'autres, l'existence de regroupement par sexe n'a pas de sens comme les rapports CW/BD où mâles et femelles forment un unique nuage de points ou encore le rapport FL/CW qui est des plus trompeurs : on peut apercevoir un semblant de clusters avec deux groupes de points et pourtant, mâles et femelles sont mélangés.

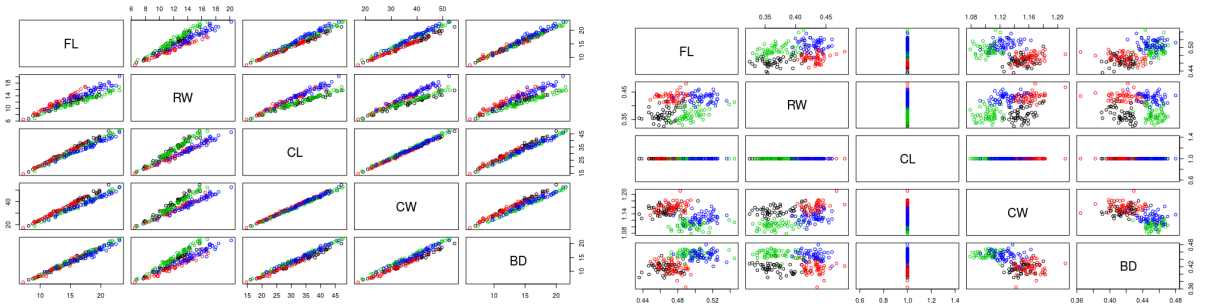


FIGURE 1.5 – Représentation des variables du dataset crabs discriminé par l'espèce et le sexe. A gauche : valeurs non transformés A droite : valeurs décorrélées après division par la variable CL

Pour finir, nous décidons d'appliquer brièvement les mêmes analyses mais en représentant 4 groupes à la fois selon le sexe et l'espèce : BM (individus B de sexe M), BF (individus

B de sexe F), OM et OF. Après transformation, des clusters sont toujours visibles mais moins distincts que lorsque nous avions discriminé avec seulement un paramètre.

1.3 Données Pima

Dans cette partie, il s'agit d'analyser le jeu de données de 532 individus de sexe féminin : nous avons à notre disposition 7 variables quantitatives que sont le nombre de grossesses, le taux plasmatique de glucose, la pression artérielle diastolique, l'épaisseur du pli cutané au niveau du triceps, l'indice de masse corporelle, la fonction de pedigree du diabète et l'âge. Nous avons en plus de ces données, une variable qualitative z qui représente la catégorie de l'individu, si il est diabétique ou non.

On s'aperçoit que les effectifs des diabétiques et des non diabétiques ne sont pas identiques, étant respectivement de 355 et 177.

En calculant les moyennes des variables classées en deux catégories on obtient :

catégories	npreg	glu	bp	skin	bmi	ped	age
1	2.93	110.02	69.91	27.29	31.43	0.45	29.22
2	4.70	143.12	74.70	32.98	35.82	0.62	36.41

Les moyennes paraissant distinctes, on pense à effectuer un test de student afin de savoir si les variables séparées en deux catégories sont indépendantes. On pourra donc savoir si le facteur « diabète » a une influence sur la valeur prise par certaines variables. Le test de student suppe l'égalité des variances sur les deux groupes que l'on compare. On effectue donc un test de variance afin de déterminer lesquels des groupes sont comparables.
En utilisant la fonction var.test, on obtient les résultats suivants : Pour l'indice de masse corporelle, on a bien une p-value=0.87. On supposera donc l'égalité. Pour l'épaisseur cutanée au niveau du triceps on a p=0.63 et pour la pression artérielle, 0.43. Les autres variables ayant des p-values proches de 0.05, on rejette l'hypothèse d'égalité des variances. Le test de student non-apparié sur les trois variables nous fournis des p-values très inférieures à 0.05 : nos deux groupes diabétique et non diabétique sont donc fortement indépendants en ce qui concerne la pression artérielle, l'indice de masse corporelle et l'épaisseur cutanée au niveau du triceps. Nous avons donc trouvés trois variables représentatives du facteur « diabète ».

2. Analyse En Composantes Principales

Cette partie a pour but de mettre en pratique la méthode d'analyse en composantes principales vue en cours. Cette méthode consiste à transformer des variables corrélées en nouvelles variables décorrélées les unes des autres. Ces nouvelles variables sont nommées « composantes principales », ou axes principaux. Elle permet de réduire le nombre de variables et de rendre l'information moins redondante, en minimisant la perte d'information.

2.1 Exercice Théorique

Dans cette exercice, nous revenons sur les données « Notes » du chapitre précédent, et plus particulièrement aux correcteurs d'examens. Le jeu de données contient les moyennes et les écarts-types par correcteur, pour le médian et le final. Les correcteurs pour lesquels des informations sont manquantes sont écartés en premier lieu.

Les données initiales sont donc :

$$X = \begin{array}{c} \\ Cor1 \\ Cor4 \\ Cor5 \\ Cor6 \\ Cor7 \\ Cor8 \end{array} \begin{array}{c} \begin{array}{cccc} moy.median & std.median & moy.final & std.final \end{array} \\ \begin{pmatrix} 10.70833 & 3.900715 & 10.94000 & 4.583303 \\ 10.23469 & 3.043268 & 13.43478 & 4.343077 \\ 10.97959 & 4.413473 & 11.82979 & 3.971743 \\ 11.50000 & 4.303584 & 13.41489 & 4.877097 \\ 10.12245 & 4.030522 & 11.90426 & 4.444878 \\ 10.74000 & 4.646056 & 11.39583 & 4.872235 \end{pmatrix} \end{array}$$

Calculer les axes factoriels de l'ACP du nuage de points défini par les quatre variables quantitatives. Quels sont les pourcentages d'inertie expliquée par chacun de ces axes ?

La première étape est de centrer la matrice précédente en colonne afin d'obtenir une matrice X_c où la moyenne de chaque colonne vaut 0. Puis, nous calculons la matrice de variance-covariance $V = \frac{1}{n} * X_c^t * X_c$. Il suffit alors de diagonaliser la matrice V pour obtenir les vecteurs et valeurs propres.

Après calcul on obtient les vecteurs propres (ou axes principaux d'inertie) suivants :

$$u_1 = \begin{pmatrix} -0.0368252505 \\ 0.2941744324 \\ -0.9550418742 \\ -0.0005681381 \end{pmatrix} \qquad u_2 = \begin{pmatrix} -0.7043052 \\ -0.6469013 \\ -0.1719624 \\ -0.2364356 \end{pmatrix}$$

$$u_3 = \begin{pmatrix} -0.23322821 \\ -0.09425237 \\ -0.02061448 \\ 0.96762396 \end{pmatrix} \qquad u_4 = \begin{pmatrix} 0.66947937 \\ -0.69720630 \\ -0.24062211 \\ 0.08832752 \end{pmatrix}$$

et les valeurs propres associées : $\lambda_1 = 0.9799$ $\lambda_2 = 0.3675$ $\lambda_3 = 0.0832$ $\lambda_4 = 0.0520$

Accompagnés des pourcentages d'inertie :

- % d'inertie expliquée par λ_1 : 66.10%
- % d'inertie expliquée par λ_2 : 24.79%
- % d'inertie expliquée par λ_3 : 5.61%
- % d'inertie expliquée par λ_4 : 3.51%

On remarque que le pourcentage d'inerties cumulé des deux premières valeurs propres atteint 91% ou 96% pour les trois premières valeurs propres, ce qui signifie que nous pouvons représenter 91% de l'information dans l'espace engendré par les deux premiers axes ou 96% pour les trois premiers axes.

Ce qui conduit aux composantes principales suivantes :

$$C = \begin{pmatrix} 1.1131294 & 0.2973223 & 0.10675046 & 0.40247613 \\ -1.5041532 & 0.8133820 & 0.01415599 & 0.06168357 \\ 0.4045437 & -0.2338454 & -0.61494553 & -0.04153965 \\ 1.1613043 & -1.0159209 & 0.11740385 & 0.08203415 \\ 0.2520651 & 0.4929044 & 0.07733933 & -0.32451159 \\ 0.8957193 & -0.3538424 & 0.29929590 & -0.18014261 \end{pmatrix}$$

Ce qui conduit aux représentations suivantes :

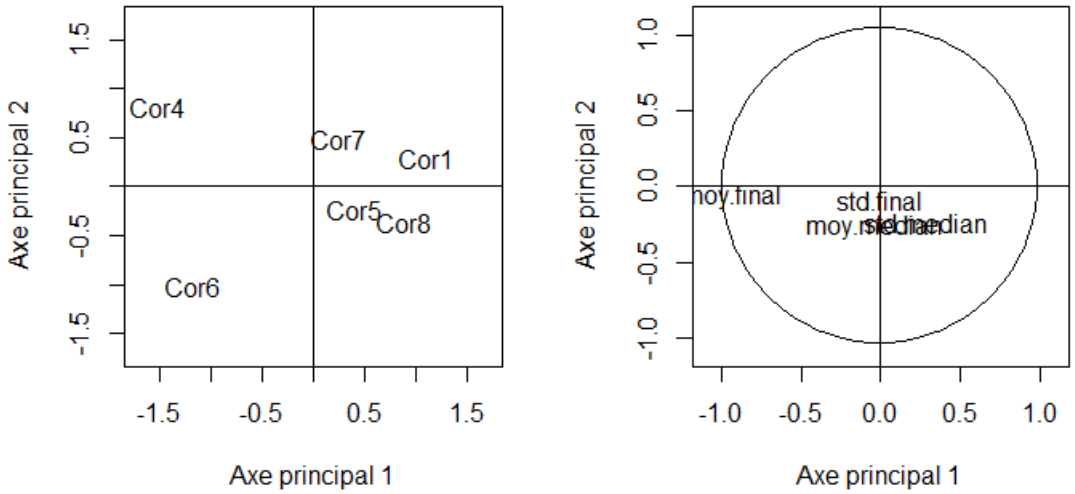


FIGURE 2.1 – A gauche : Représentation des 6 individus sur les axes 1 et 2. A droite : Représentation des 4 variables sur les axes 1 et 2

On peut par ailleurs tester la formule de reconstitution. En effet pour nos 4 variables on a l'expression $\sum_{\alpha=1}^{k} c_\alpha u'_\alpha$ pour α allant de 1 à 4 qui vaut X (la matrice des données centrée de départ).

De plus, après imputation par la moyenne des correcteurs manquants, on obtient la représentation suivante (variables : cf Annexe A.3) :

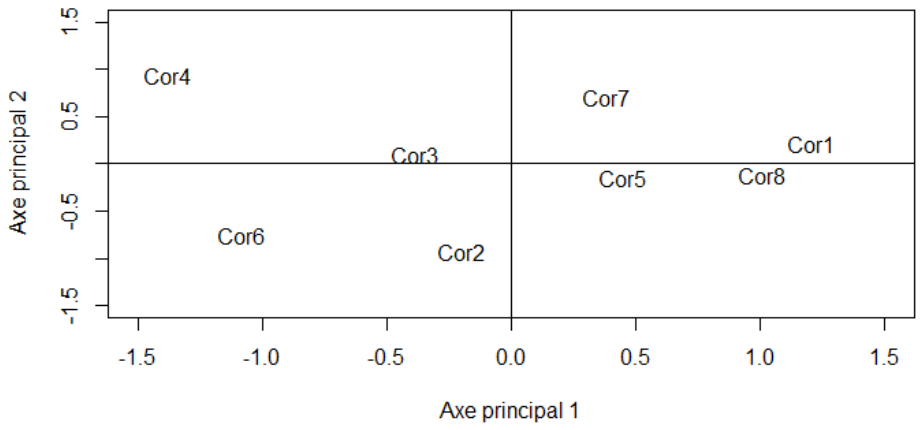


FIGURE 2.2 – Représentation des 8 individus sur les axes 1 et 2

2.2 Utilisation des outils R

On utilise le jeu de données du cours et les fonctions données dans l'énoncé du TD afin de comprendre le fonctionnement des fonctions et ce qu'elles renvoient.

- Grâce à la fonction $acp < -princomp(X)$, on calcule l'ACP (la fonction s'occupe de centrer la matrice X)
- On retrouve les valeurs propres à l'aide de $acp\$sdev^2$
- On retrouve les vecteurs propres à l'aide de $acp\$loadings$
- On retrouve la nouvelle matrice individus-variables à l'aide de $acp\$scores$

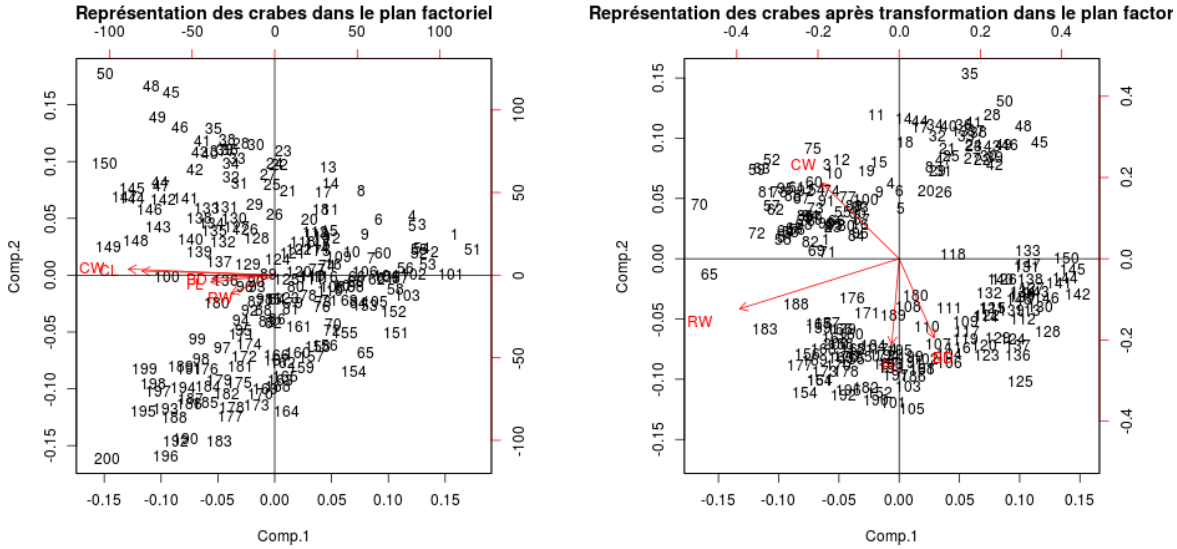
2.3 Données Crabs

Dans cette partie, nous reprenons les données crabs étudiée précédemment afin d'y appliquer une ACP. Pour celà, on utilise la fonction *princomp*() pour réaliser directement l'ACP. la fonction s'occupe du centrage de la matrice. On obtient un vecteur des 5 écarts types sur les 5 composantes.
De cette matrice, on obtient aisément les valeurs propres λ en mettant au carré ces écarts-types :

	$\lambda_1 = 140.002$	$\lambda_2 = 1.290$	$\lambda_3 = 0.995$	$\lambda_4 = 0.135$	$\lambda_5 = 0.077$
Prop. de la variance (%)	98,24	0,91	0,70	0,09	0,06
Prop. cumulée (%)	98,24	99,15	99,85	99,94	100

Nous remarquons que nous pouvons représenter plus de 98% de l'information sur le premier axe : cela est lié au faut que les variables sont très fortement corrélées comme vu auparavant et donc qu'une seule d'entre elles peut expliquer toutes les autres. Pour pallier à cela, on décorrèle les variables en divisant par CL et on recommence à faire une ACP sur les nouvelles variables en excluant CL :

	$\lambda_1 = 0.0015$	$\lambda_2 = 0.0010$	$\lambda_3 = 0.0002$	$\lambda_4 = 0.0001$
Proportion de la variance (%)	53,19	35,51	6,10	5,20
Proportion cumulée (%)	53,19	88,70	94,80	100



2.4 Données Pima

Cette partie va s'intéresser de nouveau aux données Pima étudiées avec l'ACP. On utilise de nouveaux les fonctionnalités de R présentées précédemment pour obtenir les valeurs propres :

Valeur propre	989.61	180.52	112.76	84.87	20.88	6.14	0.11
Prop. de la variance (%)	70.95	12.94	8.08	6.08	1.50	0.44	0.01
Prop. cumulée (%)	70.95	83.89	91.97	98.05	99.55	99.99	100

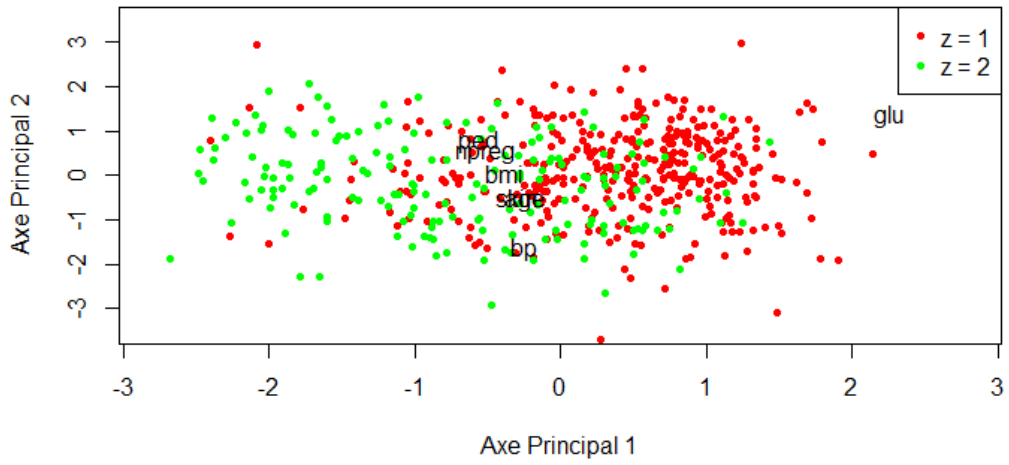


FIGURE 2.3 – Représentation des individus et des variables sur les deux premiers axes factoriels de l'ACP

On peut déjà, d'après ces résultats, critiquer les futures représentations. En effet, la proportion de variance cumulée pour les deux premiers axes n'atteint que 83.89, ce qui ne représente que peu fidèlement les variables.

On peut observer que la variable glu est bien représentée par le premier axe factoriel (c'est bien la seule), mais ce n'est pas le cas de la plupart des individus et des autres variables (C'est aussi le cas sur les autres axes cf Annexe A.4). Même si la différence n'est pas si évidente, il semble tout de même que les individus diabétiques en rouge ci-dessus aient un taux de glucose plus élevé que les individus sains. Mais cela n'est pas suffisant pour faire une réelle différenciation entre ces deux groupes, en tenant en compte de la faible représentation de l'information.

Conclusion

Nous avons pu voir que la statistique descriptive est un bon moyen d'avoir un premier aperçu sur les jeux de données que nous avons à étudier. Cependant elles ne sont parfois pas suffisamment complètes pour effectuer correctement une analyse, comme nous avons pu le constater avec le jeu de donnée « Pima ». Ainsi, l'analyse en composante principale permet de décorréler efficacement nos données afin que l'on puisse discriminer nos valeurs d'une manière pertinente.

A. Annexes

A.1 Résumé Notes SY02

```
 specialite niveau     statut                dernier.diplome.obtenu  note.median    correcteur.median   note.final     correcteur.final  note.totale
GB     :65   1: 43   Echange:  6   AUTRE 1ER CYCLE         :  6        Min.   : 0.00   Cor1:24          Min.   : 0.00   Cor1:25          Min.   : 0.900
GI     :45   2:160   UTC    :290   AUTRE 2E CYCLE          :  1        1st Qu.: 8.00   Cor2:48          1st Qu.: 9.50   Cor3:48          1st Qu.: 9.775
GM     :61   3:  5                 AUTRE DIPLOME SUPERIEUR:  4        Median :11.00   Cor4:49          Median :13.00   Cor4:46          Median :12.300
GP     :16   4: 71                 BAC                     :109       Mean   :10.92   Cor5:49          Mean   :12.38   Cor5:47          Mean   :11.845
GSM    :53   5:  7                 BTS                     :  8       3rd Qu.:14.00   Cor6:49          3rd Qu.:16.00   Cor6:47          3rd Qu.:14.800
GSU    :40   6: 10                 CPGE                    : 41       Max.   :20.00   Cor7:49          Max.   :19.50   Cor7:47          Max.   :18.400
HuTech: 1                          DEUG                    :  3       NA's   :3       Cor8:25          NA's   :12      Cor8:24          NA's   :12
ISS    : 4                         DUT                     : 92                       NA's: 3                          NA's:12
TC     :11                         ETRANGER SECONDAIRE     :  5
                                   ETRANGER SUPERIEUR      : 11
                                   INGENIEUR               :  1
                                   LICENCE                 :  9
                                   NA's                    :  6

resultat
F    :49
Fx   :34
E    :37
D    :44
C    :58
B    :42
A    :20
NA's:12
```

FIGURE A.1 – Résumé des caractéristiques des différentes variables du jeu de données Notes

A.2 Comparaisons des notes données par les correcteurs aux examens

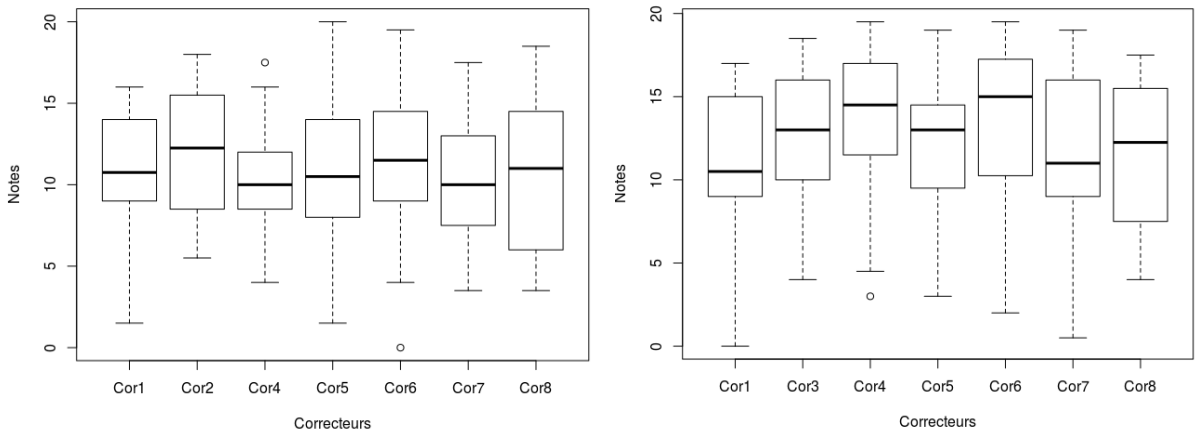


FIGURE A.2 – A gauche : Diagrammes en boites des notes du médian par correcteur. A droite : Diagrammes en boites des notes du final par correcteur

A.3 Représentation des Variables ACP Correcteurs

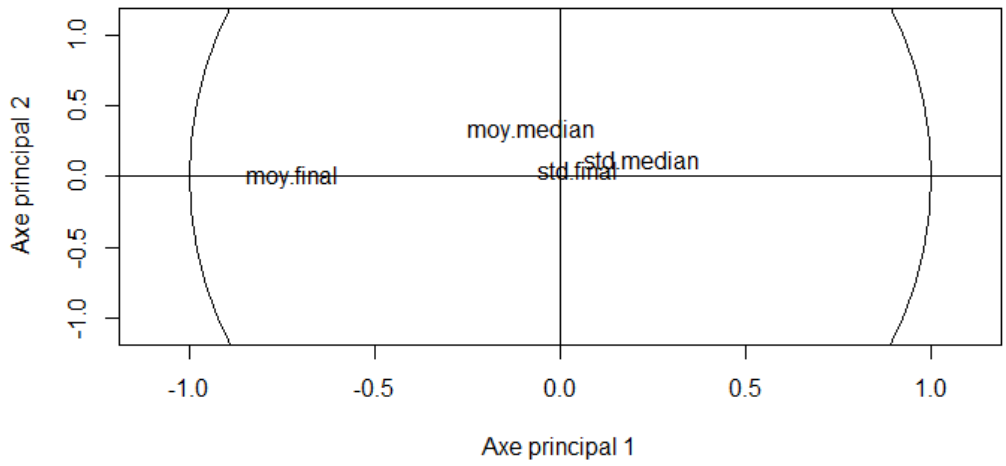


FIGURE A.3 – Représentation des 4 Valeurs avec ACP après imputation par la moyenne

A.4 Variables dans le repères composé des vecteurs Propres après ACP et centrage pour Pima

```
         [,1]        [,2]       [,3]       [,4]        [,5]
npreg -0.5159274  0.49589166  0.3637487  0.2193017  0.16643583
glu    2.2502114  1.37359648  0.1567246 -0.3118077  0.19344275
bp    -0.2507970 -1.64192942  0.7096587 -1.6694080  0.41117671
skin  -0.3008539 -0.46386966 -1.7420215  0.5232875  1.23732800
bmi   -0.3833052  0.04896506 -0.8524571 -0.2211796 -2.05095597
ped   -0.5586767  0.77787688  0.1131979 -0.1864738  0.13855851
age   -0.2406512 -0.59053101  1.2511487  1.6462798 -0.09598583
             [,6]       [,7]
npreg  2.2188364 -0.3630605
glu   -0.2716732 -0.3801716
bp    -0.2766813 -0.3701812
skin  -0.3166562 -0.3776436
bmi   -0.2246676 -0.3933300
ped   -0.3017477  2.2676825
age   -0.8274104 -0.3832955
```