SY09 TP1
Statistique descriptive, Analyse en composantes principales

Thibault Subreville & Yuting Chen

P17

L'objectif de ce TP est d'apprendre à manipuler une grande quantité de données en utilisant des méthodes exploratoires élémentaires et l'analyse en composantes principales. L'outil statistique utilisé est le logiciel R.

1 Statistique descriptive

1.1 Notes

Dans cette partie nous utilisons le tableau de données *sy02-p2016.csv* qui est constitué de 296 étudiants décrits par 11 variables dont 4 sont quantitatives (**niveau**: le semestre d'étude, **note.median**: la note du médian, **note.final**: la note du final, **note.total**: la note de final pour l'UV) et 7 sont qualitatives (**nom**: le nom de l'étudiant anonymisé, **spécialité**: la branche d'étude, **statut**: l'établissement d'origine, **niveau**: le semestre d'étude, le **dernier diplome obtenu**, le **correcteur du median** et le **correcteur du final**, et enfin le **resultat** de l'UV). Des valeurs sont manquantes, comme par exemple des notes de médian ou de final pour un étudiant. Nous supposons que ces valeurs sont le fruit de l'absence de l'étudiant à l'épreuve.
Nous cherchons en particulier à étudier les liens statistiques entre les variables la note obtenu à l'UV et la branche, le correcteur ou encore avec la note du médian.
Sans analyser les données, nous pouvons supposer que les notes d'un étudiant sont corrélés, i.e. une note similaire entre le médian et le final. Nous pouvons aussi présuposé qu'un correcteur note plus durement qu'un autre.

GB	GI	GM	GP	GSM	GSU	HuTech	ISS	TC
61	43	58	16	52	40	1	4	9

Résumé de la répartition des étudiants par branche

Après le traitement préliminaire qui consistait à enlever de l'étude tous les étudiants auquel il manquait la valeur du final, nous arrivons à un échantillon total de 284 individus.

- **Le lien entre la note de l'UV et la branche**
  En ce qui concerne la note des étudiants, on constate que ces dernières possèdent des dispersions similaires indépendamment de la branche d'étude. Le jeu de données conernant l'unique étudiant en HutTech n'étant pas significatif, sa valeur y est ici juste informel. A la même image des TC et des ISS qui à eux deux forment une population de 13 individus, il nous est donc impossible de déduire des résultat statistiques significiatfs dessus.

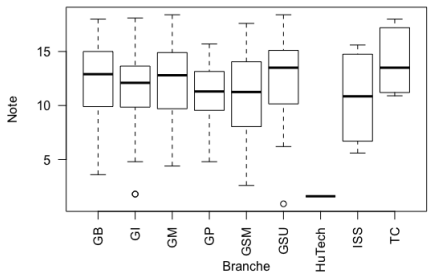

Figure 1: Diagramme en boîte de la note des étudiants en fonction de la branche d'étude

Nous réalisons un test du Chi2 à 1 degré de liberté avec la fonction **chisq.text**. Pour mettre en évidence, un lien significatif entre les résultats d'un étudiant provenant de DUT et d'un étudiant n'en provenant pas. Nous considérons l'hypothèse nulle (H_0): les deux variables sont indépendantes. Pour cela, nous nous sommes intéréssés au critère d'obtentoin de l'UV:

	DUT	non DUT
Réussite à l'UV	57	163
Echec à l'UV	32	46

Résumé de la répartition des étudiants en fonction de la réussite à l'UV

On obtient une **p-value** $= 0.018$ à 10^{-3} près. On conclut de ce test qu'au seuil de 5%, on peut rejeter l'hypothèse initial d'indépendance des 2 variables.
Cet écart par rapport aux autres filières peut vraisemblalement s'expliquer par le faible nombre d'heure en mathématiques que la filière DUT à reçu durant les deux années précédents leurs arrivés à l'UTC.

- **Le lien entre les notes et le correcteur**
  Nous pouvons observer que les notes moyennes évaluées par 8 correcteurs ne sont pas très différentes, il est juste pour chaque étudiant. Selon la supposition au départ, nous pouvons alors rejeter l'hypothèse d'un correcteur notant plus durement que les autes.

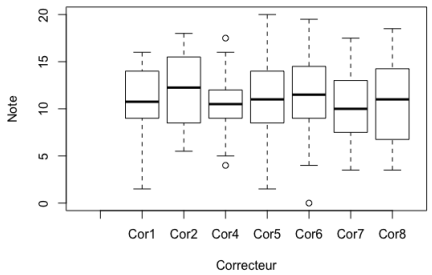

Figure 2: Diagramme en boîte des correcteurs et des notes au final de SY02

- **Le lien entre la note du médian, du final et de l'UV**
  Les résulats entre le final/median et la note totale sont corrélés par définition du calcul de la note total; ce résulat ne nous étonne guère. En revanche ce qu'il l'est un peu plus, est la

corrélation modéré (r=0.43) entre le final et le médian de SY02. La moyenne des examens sont respectivement pour le médian de **11.05** et pour le final de **12.38**. On constate donc

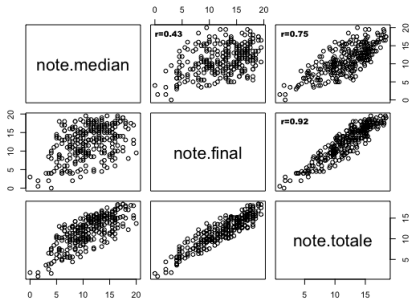

Figure 3: Graphique matriciel des notes obtenues par les étudiants

une volatilité des notes significatives pour une même personne en général entre son médian et son final. Cela peut s'expliquer en partie par une moyenne différente entre le médian et le final.

- **Conclusion sur l'étude des notes de SY02**
  Après avoir analysé les données, on peut déterminer les liens entres les notes, les branches et les correcteur. On en déduit qu'il n'y a pas de lien significative entre les différents correcteurs. Concernant les branches, il est difficile de se prononcer car nous ne disposons pas d'individus dans chaque filières. En revanche, si on enlève les TC, HuTech et ISS on n'observe aucune prédominance d'une branche à une autre. Enfin, le dernier dîplome obtenu et donc le cursus de l'étudiant à une influence sur les notes et l'obtention de l'UV comme nous avons pu le voir avec le cas du cursus DUT.

1.2 Données crabs

Le jeu de données considéré est constitué de 200 crabes décrits par huit variables dont trois variables qualitatives (sexe, espèce et l'index) et cinq quantitatives correspondant à différentes longueurs morphologiques mesurées en mm sur les crabes.

- **Différences de caractéristiques morphologiques selon l'espèce et le sexe**

  Pour déterminer s'il existe ou non des différentes morphologiques significatives entre le sexe des crabes ou leur espèce, une analyse statistique a été réalisé en comparant deux à deux les variables quantitives.

  Nous remarquons qu'il est presque impossible de distinguer les espèces entre elles. On obtient de résultas similaires mais moins variant en fonction de leur sexe.

- **Corrélation des variables et identification de la cause**

	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	0.99	0.98
CW	0.96	0.90	0.99	1.00	0.97
BD	0.99	0.90	0.98	0.97	1.00

Table de corrélation des variable quantitatives décrivant les crabes à 10^{-2} *près*

Cette table nous montre une forte corrélation entre toutes les variables deux à deux (coefficient de corrélation supèrieur à 0.89). Cette corrélation peut vraisemblablement s'expliquer

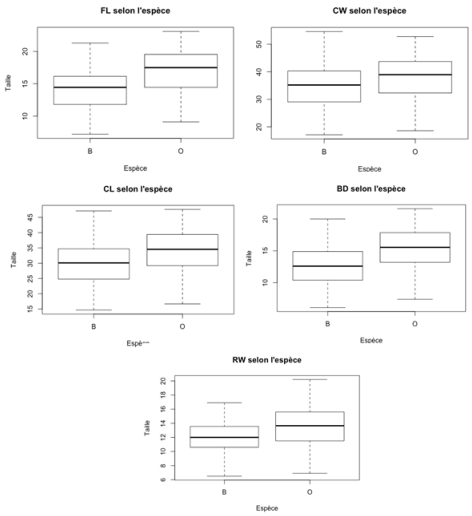

Figure 4: Diagramme en boîte des espèces de crabes en fonction de leurs mesures morpholiques

par *l'effet de taille* des crabes, i.e. un crabe avec un grande carapace n'a pas un petit corps. Pour s'affranchir de ce phénomène, on divisera les valeurs du tableau initial d'un crabe par la longueur de sa carapace, CL. CL étant le paramètre ayant la corrélation la plus grande avec les autres paramètres. On s'affranchit alors de l'effet taille, on obtient les matrices de corrélations suivantes:
L'espèce B étant en bleu et l'espèce O en orange. Le premier constat qu'on en tire est que

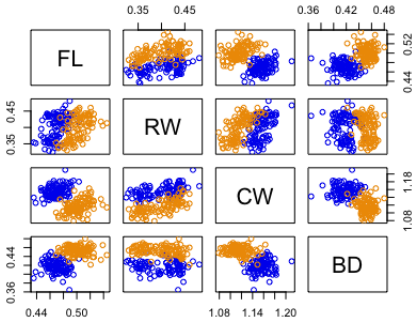

Figure 5: Graphique matriciel des espèces de crabes en fonction de leurs morphologies

les espèces sont maintenant beaucoup plus différenciables, ce qui permet des les identifier par rapport à l'espèce. Nous obtenons un graphe matriciel similaire en fonction du sexe.

1.3 Données Pima

Le jeu de données considéré est constitué de 532 individus de sexe féminin décrits par huit variables (dont une qualitative) : le nombre de non-diabétiques est 355, le nombre de diabétiques est de 177.

- **Différents influences selon le diabète**
  D'après les diagrammes en boites on constate que pour une diabétique les variables npreg, glucose, nombre de grossesse, masse corporelle, pedigree, et l'âge sont élévées.

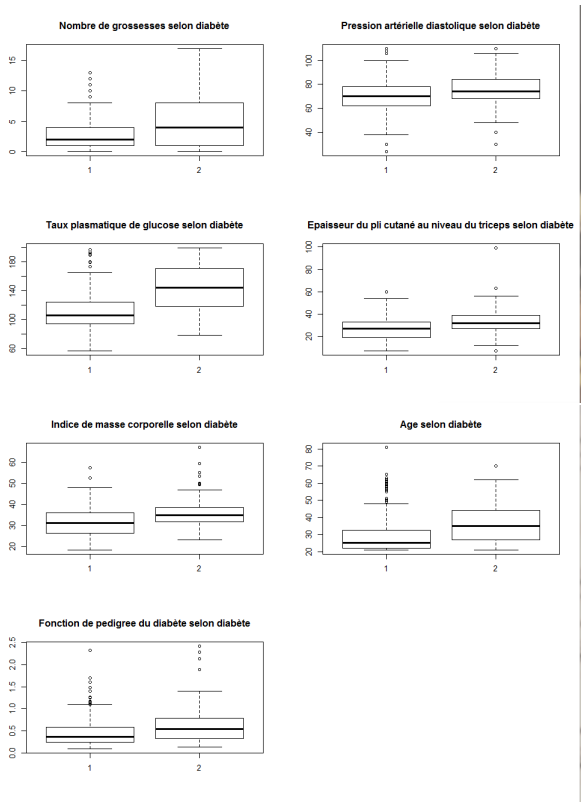

Figure 6: Diagramme en boîte sur les différents effets sur le diabète

- **Corrélation des variables**

  On peut également voir sur le figure 7 qu'il existe une forte corrélation (r=0.65) entre l'épaisseur du pli cutané au niveau du triceps (skin) et l'indice de masse corporelle (bmi).

	npreg	**glu**	**bp**	**skin**	**bmi**	**ped**	**age**
npreg	1.00	0.13	0.20	0.09	0.01	0.01	0.64
glu	0.13	1.00	0.20	0.23	0.25	0.17	0.28
bp	0.20	0.22	1.00	0.23	0.31	0.01	0.35
skin	0.09	0.23	0.23	1.00	0.65	0.12	0.16
bmi	0.01	0.25	0.31	0.65	1.00	0.15	0.07
ped	0.01	0.17	0.01	0.12	0.15	1.00	0.07
age	0.64	0.28	0.35	0.16	0.07	0.07	1.00

Table de corrélation des variables quantitatives décrivant les données Pima à 10^{-2} *près*

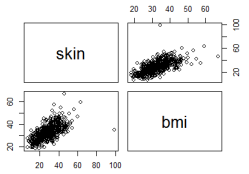

Figure 7: Graphique matriciel de la masse corporelle et du pli cutané

On remarque aussi le nombre de grosseses et l'âge sont corrélés mais cela correspond plus à une évidence universelle. De ce fait, cette corrélation n'est pas pertinente pour l'étude.

- **Conclusion sur l'étude Pima**

  Après avoir analysé les données, on peut dire que le diabète est une sorte de maladie lié à l'âge, au mode de vie et à l'obésité. Les personnes obèses, agées sont plus susceptible d'avoir le diabète. L'hérédité (pedigree) y joue aussi un rôle.

2 Analyse en composantes principales

2.1 Exercice théorique

Dans un premier temps, on centre la matrice M donnée grâce à la fonction **scale** (on centre mais on ne réduit pas la matrice M). Ensuite, on calcule sa matrice de variance grâce à la formule $\frac{1}{n}X^tX$ où $n = 6$. Nous avons dû supprimer deux correcteurs car ils n'avaient pas participer soit à la correction du médian soit à la correction du final. La fonction **eigen** nous permet d'obtenir les valeurs propres et les vecteurs propres associés, les axes factoriels sont:

$$\begin{pmatrix} -0.093 \\ 0.274 \\ -0.957 \\ 0.002 \end{pmatrix} \quad \begin{pmatrix} -0.659 \\ -0.710 \\ -0.140 \\ -0.207 \end{pmatrix} \quad \begin{pmatrix} -0.480 \\ 0.178 \\ 0.099 \\ 0.853 \end{pmatrix} \quad \begin{pmatrix} 0.572 \\ -0.624 \\ -0.234 \\ 0.479 \end{pmatrix}$$

associés respectivement aux valeurs propres: 0.980, 0.327, 0.105 et 0.037 à 10^{-3} près.
Concernant le pourcentage d'inertie de ces axes, nous avons:

Vecteur propre	**1er**	**2e**	**3e**	**4e**
Pourcentage d'inertie des axes principaux	67.638%	22.573%	7.233 %	2.556%
Pourcentage d'inertie sous-espaces principaux	67.638%	90.211%	97.444%	100%

Pourcentage d'inertie des axes factorielles à 10^{-3} près

On obtient la matrice C=XU suivante à 10^{-3} près :

$$\begin{pmatrix} 1.155 & 0.280 & -0.015 & 0.286 \\ -1.487 & 0.817 & -0.055 & 0.096 \\ 0.342 & -0.322 & -0.657 & -0.104 \\ -1.176 & -0.933 & 0.171 & 0.050 \\ 0.304 & 0.417 & 0.209 & -0.342 \\ 0.861 & -0.259 & 0.346 & 0.012 \end{pmatrix}$$

On calcule la somme suivante pour k=4: $\sum_{\alpha=1}^{4} C_\alpha U_\alpha^t$, et on ré-obtient la matrice X de départ. En effet, on retrouve que la meilleur projection dans la dimension de départ de X (soit 4) est bien elle-même. Pour k<4, on réalisé une projection dans un espace de dimension inférieur à celui de base, on perd donc des informations.

2.2 Utilisation des outils R

L'objectif de cette exercice est de faire une ACP sur les données des notes de cours en utilisant une fonction plus directe que ceux utilisés précédemment. La fonction **bliplot.princomp** permet d'accéder à des options graphiques supplémentaires telle que les fonctions scale et pc.biplot qui permettent de rédimensionner l'affichage des variables dans l'ACP.
La fonction suivante acp = **princomp(notes)** permet de réaliser une ACP sur la matrice notes.
Les résultats peuvent être retrouvées de la manière suivante :
Valeurs propres : $acp\$sdev$
Matrice diagonale des valeurs propres D : $D = diag((acp\$sdev)^2)$
Matrice des axes principaux U : $U = acp\$loadings$
Matrices des composantes principales C : $C = acp\$scores$ ou $C = XU$
La fonction **summary** permet d'obtenir un aperçu des éléments de statistiques descriptifs des variables et des individus (variance, moyenne, quartile...).

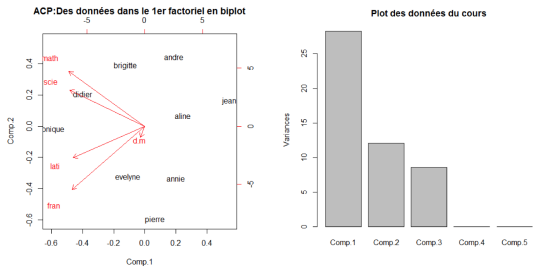

Figure 8: Illustration des fonctions plot et biplot en ACP

2.3 Données Crabs

- **ACP sans traitement**

  Sans traitement préalable, l'utilisation de l'ACP ne permet pas de distinguer les crabs selon leur espèce et/ou leur sexe. Ce phénomène s'explique par l'effet taille que nous avons exposé à la section précédente.
  Légende: vert: Bleu+Male, rose: Bleu+Femelle, bleu: Orange+Male, rouge: Orange+Femelle

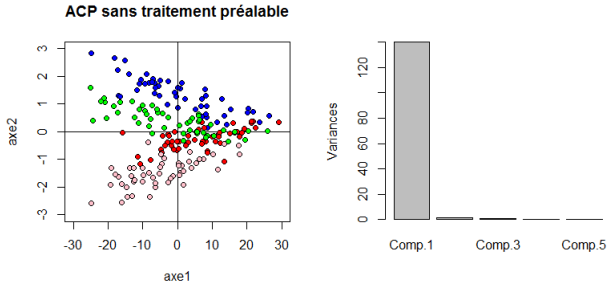

Figure 9: Utilisation de l'ACP sans traitement préalable

- **ACP avec traitement**

  Il est possible d'améliorer la représentation précédente en s'affranchissant de l'effet de taille. Pour cela nous utilisons la même méthode que dans la section précédente. La

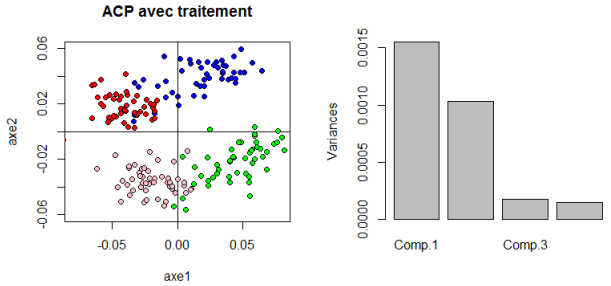

Figure 10: Utilisation de l'ACP avec traitement

distinction entre les espèces est plus nette, ainsi que la distinction entre les sexes au sein d'une même espèce.

2.4 Données Pima

Dans le cas des données Pima, en réalisant une ACP on obtient la figure 11. Cette ACP n'est pas concluant car on ne peut distinguer clairement la population diabétique (en rouge) et la population non-diabétique (en noir).

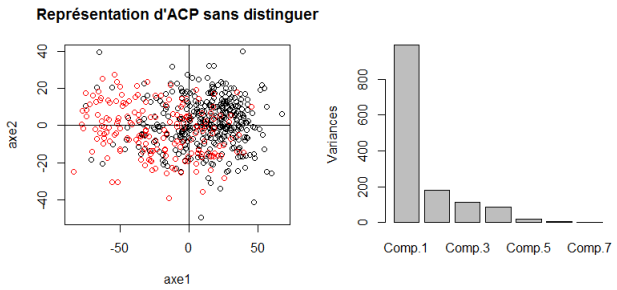

Figure 11: ACP des données Pima non concluante

Il paraît alors peu probable de trouver une représentation simple de ces données sans perdre beaucoup d'informations sur les variables qui représentent chaque individu. Nous arrivons à des représentations semblables dans le deuxième plan principal ou même dans le plan de la deuxième et troisième composante. L'ignorance d'une variable donne aussi un résultat similaire...
On peut alors poser l'hypothèse qu'il nous manque des variables pour décrire un individu comme son taux d'insuline dans le sang.
On est typiquement confronté au problème des dimensions et de la corrélation entre les variables. L'ACP n'est pas toujours la solution à tous les problèmes de réduction des dimensions.

3 Conclusion du TP

Ce TP nous permis de voir différents modes de visualisation des données statistiques. L'ACP nous a montré tant bien ces points forts en nous permettant réduire le nombre de variables et de rendre l'information moins redondante. Mais, nous avons aussi pu apercevoir les limites de l'ACP qui nécessite parfois un traitement des données pour obtenir une représentation pertinente ou qui même avec un traitement ne donne pas une représentation concluante.