SY09 - TP01
Statistique descriptive, Analyse en composantes principales

Cristian GARRIDO

Avril 24, 2014

OBJETIF

L'objectif de ce TP est de manipuler des données afin de mettre en évidence les principaux éléments qui ont une influence sur des variables (temps de gestation, poids des bébés, taille de crabes). La méthode de l'ACP nous aidera dans l'analyse des tableaux de données.

1 STATISTIQUE DESCRIPTIVE

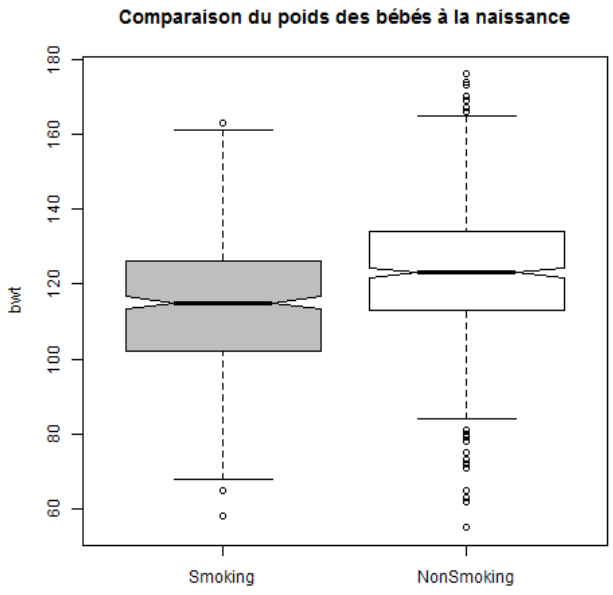
1.1 DONNES BABIES

Dans ce exercice, nous disposons d'un jeu de données babies constitué de 1236 bébés décrits par 23 variables. Dans notre tp, nous n'utilisons que 8 variables : 5 quantitatives (le poids à la naissance, la durée de gestation, le nombre de grossesses précédentes, la taille de la mère et le poids de la mère) et 3 qualitatives (l''âge de la mère, si la mère fume ou non et le niveau d'éducation de la mère).

1.1.1 Quelle est la différence de poids entre les bébés nés de mères qui fumaient durant leur grossesse et celles qui ne fumaient pas ?

En suite, nous avons le résumé des les valeures statistiques generales des poids à la naissance sur les meres qui fument ou non :

	Min	1er Quartile	Médiane	Moyenne	3e Quartile	Max
Mère fumeuse	55	113	123	123	134	176
Mère non fumeuse	58.0	102.0	115.0	114.1	126.0	163.0

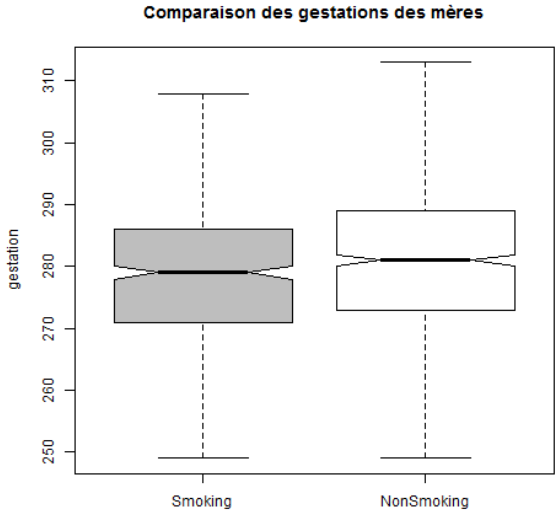


Nous pouvons remarquer une différence importante entre le poids moyens des bébés des mères fumeuses et non fumeuses (poids moyen non fumeuse > poids moyen fumeuse).
Nous représentons ces résultats par deux boites à moustaches.
Nous remarquons la présence de plus de valeurs atypiques pour les bébés dont les mères ne fument pas. La différence de poids entre les bébés nés de mères fumeuses ou non fumeuses est visible sur ces diagrammes.
Afin de déterminer si cette différence est significative, nous représentons les intervalles de confiances à 95%. Nous remarquons que ces deux intervalles sont disjoints. Notre hypothèse est donc vérifiée : il y a bien une différence de poids significative entre les bébés nés de mères fumeuse et ceux nés de mères non fumeuses.

Est-ce qu'une mère qui fume durant sa grossesse est enclin à avoir un temps de gestation plus court qu'une mère qui ne fume pas?

On remarque ici que le temps de gestation est plus élevé chez les mères non fumeuses que chez les mères fumeuses, cette différence reste tout de même très légère. Grâce au test de Student, on obtient les résultats suivants :

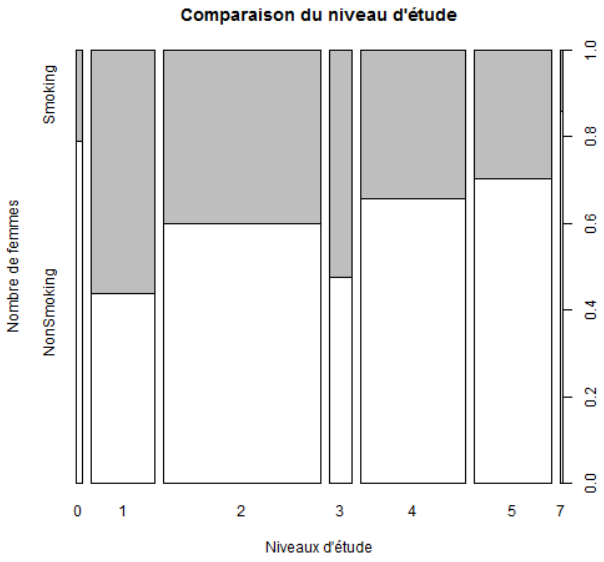
	Min	1er Quartile	Médiane	Moyenne	3e Quartile	Max
Mère fumeuse	223	271	279	278	286	330
Mère non fumeuse	148.0	273	281.0	280.2	289.0	353.0



Nous pouvons remarquer que les médianes sont légèrement différentes : le temps de gestation médian d'une mère fumeuse est plus faible que le temps de gestation médian d'une mère non fumeuse.

Les diagrammes en moustaches nous montrent que les distributions sont très proches.
Si nous zoomons sur les intervalles de confiance à 95%, nous pouvons constater que ceux-ci se chevauchent. Notre hypothèse sur l'influence de la caractéristique fumer ou non sur le temps de gestation est donc rejetée.

1.1.2 Le niveau d'étude a-t-il une influence sur le fait que la mère soit fumeuse?



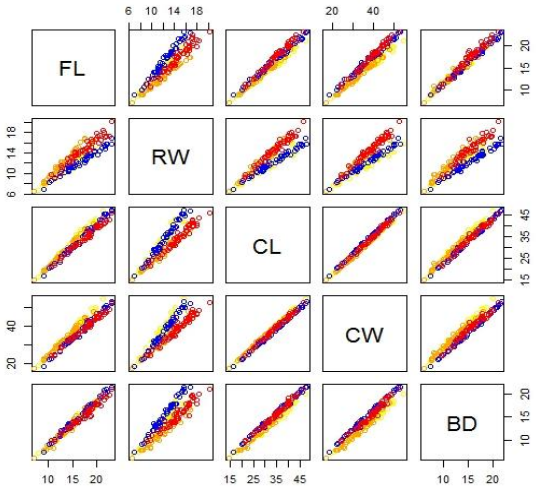
Comme le niveau d'étude est une variable qualitative, nous ne pouvons pas faire d'analyse numérique. Nous allons donc représenter la fréquence de femme fumant ou non par niveau d'éducation.
A travers ce graphique on peut observer une différence entre les femmes fumeuses et celles qui ne fument pas en fonction de leurs niveau d'étude. La proportion de femmes fumant semble diminuer généralement plus le niveau d'étude est élevé.
Afin de vérifier cette hypothèse, nous allons effectuer un test du chi2. Notre hypothèse HO est : le niveau d'étude n'a pas d'influence sur le fait que la mère soit fum0euse ou non. Nus effectuons le test du chi2 en utilisant la fonction chisq.test sur R sur les variables qualitatives niveau d'étude et femme fumeuse ou non fumeuse. Le résultat est le suivant : le degré de liberté est égal à 6, la valeur du test est 42,5 et p-value = 1.459e-07. Comme p-value est inferieur à 0.05 alors l'hypothèse Ho est rejetée. Il existe donc bien un lien entre le niveau d'étude d'une mère et le fait qu'elle soit fumeuse ou non.

1.2 DONNÉES CRABS

Le jeu de données proposé est constitué de 200 crabes décrits par sept variables dont cinq sont quantitatives.

Ici nous avons un graph multi dimensionnel représente la distribution des crabes en fonction de leur caractéristique morphologiques (deux à deux). Au vue de ce graph, il semble impossible de distinguer les crabes en fonction de leur sexe ou de leur espèce.

Rouge = Femelle Orange
Bleu = Male Orange
Orange = Femelle Blanche
Yellow = Male Blanc



Nous pouvons remarquer que chaque graph a la même allure : les points sont repartis selon une droite parallèle à chaque autre graph. Nous pouvons conclure que les caractéristiques morphologiques semblent être toutes fortement corrélées selon un coefficient commun. Afin de vérifier cette intuition, observons le tableau de corrélation suivant :

	FL	RW	CL	CW	BD
FL	1	0.9069876	0.9788418	0.9649558	0.9876272
RW	0.9069876	1	0.8927430	0.9004021	0.8892054
CL	0.9788418	0.88927430	1	0.9950225	0.9832038
CW	0.9649558	0.9004021	0.9950225	1	0.9678117
BD	0.9876272	0.8892054	0.9832038	0.9678117	1

Nous pouvons remarquer que tous les coefficients de corrélations sont compris entre 0,88 et 0,99. Toutes les caractéristiques sont donc toutes fortement corrélées. Cette forte corrélation est vraisemblablement due à « l'effet taille » influençant les valeurs de chaque données morphologiques. Cela biaise les données et ne permet de distinguer que les gros crabes des petits et non de les distinguer en fonction de leur sexe ou de leur espèce.

Afin de pouvoir évaluer une possible corrélation entre les différentes variables, nous allons donc normaliser ces variables en divisant les données par la somme des données pour chaque individu.

Après le traitement des données, nous obtenons ainsi le matrice de corrélation suivante :

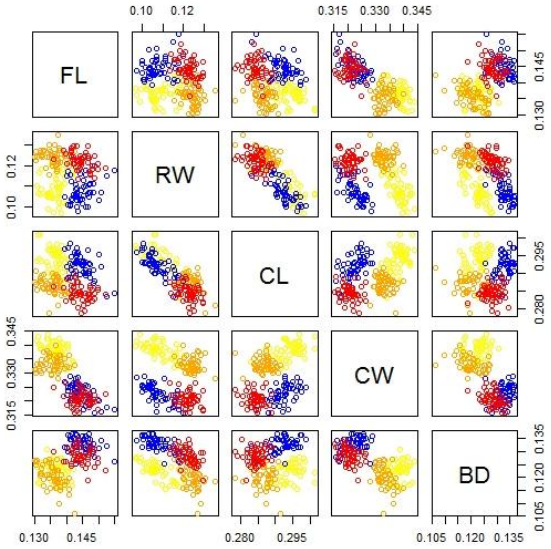
	FL	RW	CL	CW	BD
FL	1	-0.1183474	-0.24026563	-0.7825549	0.51974244
RW	-0.1183474	1	-0.82661717	-0.2020264	-0.46394103
CL	-0.2402656	-0.8266172	1	0.4186543	0.09676383
CW	-0.7825549	-0.2020264	0.41865431	1	-0.64982922
BD	0.5197424	-0.4639410	0.09676383	-0.6498292	1

Au vue de ces coefficients de corrélations, nous pouvons en déduire que les variable RW et CL sont corrélées (0.8266172).

La normalisation des données nous permet d'avoir une meilleure représentation des données suivant le sexe et l'espèce du crabe pour chaque caractéristique morphologique.

Nous pouvons maintenant distinguer des classes de sexes et des classes d'espèces.

Ainsi si l'on voulait déterminer le sexe et l'espèce d'un crabe, il faudrait placer un point pour chacune des valeurs des données morphologiques sur ce graph. Nous pourrions ainsi déterminer à quel sexe et à quelle espèce il appartient.



2 L'ANALYSE EN COMPOSANTES PRINCIPALES (ACP)

2.1. EXERCICE THÉORIQUE

2.1.1 Composantes principales – Représentation dans le premier plan factoriel

Trois variables mesurées sur quatre individus fournissent le tableau suivant :

$$A = \begin{pmatrix} 3 & 4 & 3 \\ 1 & 4 & 3 \\ 2 & 3 & 6 \\ 2 & 1 & 4 \end{pmatrix} \qquad X = \begin{pmatrix} 1 & 1 & -1 \\ -1 & 1 & -1 \\ 0 & 0 & 2 \\ 0 & -2 & 0 \end{pmatrix}$$

Les moyennes des trois variables sont respectivement 2, 3 et 4. Le tableau centré en colonne X est obtenu en soustrayant à chaque colonne la moyenne correspondante. On peut aussi utiliser la fonction scale.

$$\text{Matrice de variance -> = } S = X^t D_p X = \frac{1}{4} X^t X = \begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 1.5 & -0.5 \\ 0 & -0.5 & 1.5 \end{pmatrix}$$

La diagonalisation de la matrice de variance (grâce à la fonction eigen()) fournit les valeurs propres suivantes ainsi que les vecteurs propres normés ou axes principaux :

$$\lambda_1 = 0.5 \quad \lambda_2 = 1 \quad \lambda_3 = 2 \quad \mu_1 = \begin{pmatrix} 0 \\ -0.707 \\ 0.707 \end{pmatrix} \quad \mu_2 = \begin{pmatrix} 0 \\ 0.707 \\ 0.707 \end{pmatrix} \quad \mu_3 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

Pour choisir le nombre d'axes à retenir, on s'appuie généralement sur les pourcentages d'inertie expliquée par les différents sous-espaces.

$$\mu_1 = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} * 100 = \frac{0.5}{\text{trace}(\lambda)} * 100 = 57{,}14\%$$

$$\mu_2 = \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} * 100 = \frac{0.5}{\text{trace}(\lambda)} * 100 = 85{,}71\%$$

$$\mu_3 = \frac{\lambda_1 + \lambda_2 + \lambda_3}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{0.5}{\text{trace}(\lambda)} * 100 = 100\%$$

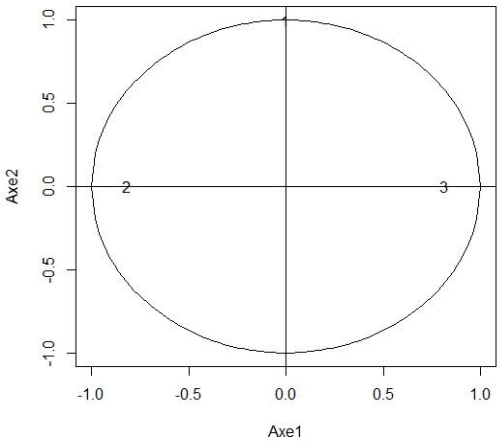
Nous remarquons que le premier plan factoriel permet de représenter 85,71% de l'information totale. Ce plan est composé des vecteurs μ_1 et μ_2.
La matrice des composantes principales $C = XMU = XU$ est la suivante :

$$\begin{pmatrix} -1.414 & 0 & 1 \\ -1.414 & 0 & -1 \\ 1.414 & -1.414 & 0 \\ 1.414 & 1.414 & 0 \end{pmatrix}$$

2.1.2 Représentation des variables dans le premier plan factoriel

La représentation des 4 individus dans le premier plan factoriel nous montre que les deux premiers individus sont confondus. Ceci n'est pas surprenant car la coordonnée les distinguant dépend du troisième axe factoriel.
Afin de tracer la représentation des trois variables dans le premier plan factoriel, nous calculons les corrélations entre ces variables.

$$D = cor(X, C) = \begin{pmatrix} 0 & 0 & 1 \\ -0.816 & 0{,}577 & 0 \\ 0.816 & 0.577 & 0 \end{pmatrix}$$



2.1.3 Calcul de l'expression $\sum_{\alpha}^{k} \boldsymbol{c_\alpha u'_\alpha}$ pour les valeurs k=1,2 et 3

K=1	K=2	K=3
$\boldsymbol{X} = \begin{pmatrix} \mathbf{0} & \mathbf{1} & \mathbf{-1} \\ \mathbf{0} & \mathbf{1} & \mathbf{-1} \\ \mathbf{0} & \mathbf{-1} & \mathbf{1} \\ \mathbf{0} & \mathbf{-1} & \mathbf{1} \end{pmatrix}$	$X = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 2 \\ 0 & -2 & 0 \end{pmatrix}$	$X = \begin{pmatrix} 1 & 1 & -1 \\ -1 & 1 & -1 \\ 0 & 0 & 2 \\ 0 & -2 & 0 \end{pmatrix}$

On retrouve bien le résultat X = CU' quand k = 3

2.2 UTILISATION DES OUTILS R.

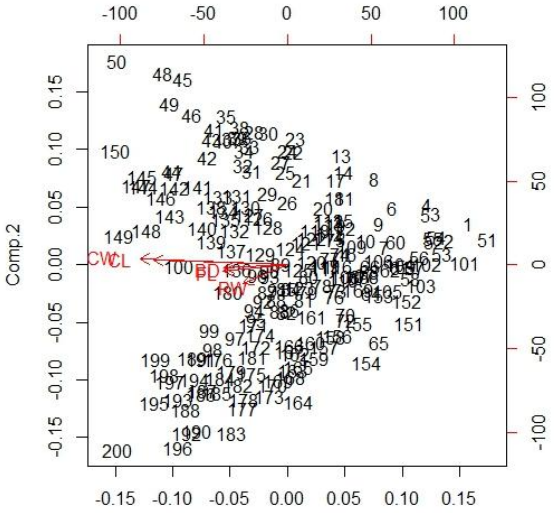
En utilisant les fonctions de R, on retrouve facilement les résultats de l'ACP:

- ACP <- princomp(notes) réalise l'ACP et nous retourne l'écart-type de chaque composante principale
- ACP$loadings retourne les axes factoriels (vecteurs propres de la matrice de variance)
- APC$scores retourne les composantes principales
- biplot(ACP) projette les données et les caractères dans le plan principal formé des deux axes factoriels de poids maximal.
- plot(ACP) représente l'histogramme des variances des composantes principales.

plot (ACP) nous permet de constater graphiquement que quasiment toutes les données sont représentées sur le premier plan factoriel. En effet, la variance cumulée de ces deux plans atteint plus de 90%. Ceci est suffisant pour s'assurer que notre analyse de données sera solide.
Biplot(ACP) représente tous les individus et les variables (vecteurs rouges) dans le premier plan factoriel. Ceci nous permet d'évaluer graphiquement les corrélations entre variables (angle formé par les deux vecteurs): les vecteurs de deux variables indépendantes forment un angle de 90°.

2.3 TRAITEMENT DES DONNÉES CRABS

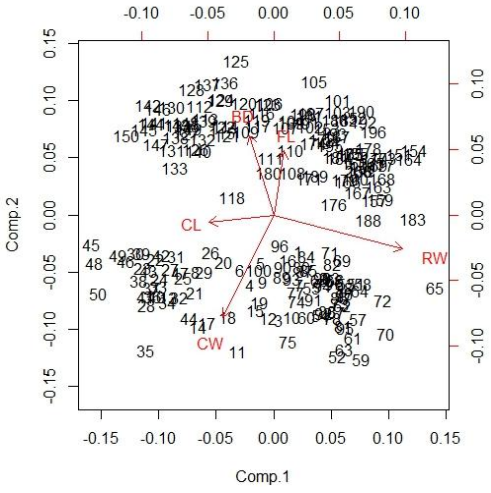
2.3.1 Tester tout d'abord l'ACP sur **crabsquant,** sans traitement préalable. Que constatez-vous?



Dans cette partie, nous avons travaillé directement avec l'ACP en utilisant la fonction princomp. Cette fonction permet d'effectuer une ACP et renvoie un objet dont la structure est composé des vecteurs et matrices utiles lors de la réalisation d'une ACP (écarts types sur chacun des axes factoriels, axes factoriels, composantes principales). Nous pouvons appliquer à cet objet la fonction biplot nous permettant d'obtenir une représentation graphique des 200 individus et la projection des 5 variables dans le plan principal. Ce graph ne nous permet pas de distinguer différents types de crabes. Nous pouvons cependant remarquer que les 5 vecteurs représentant les 5 variables ne dépendent quasiment que de la composante 1. Si l'on regarde de plus près les pourcentages d'inerties expliquées, on se rend compte que l'axe 1 représente la majorité des informations (98 %).

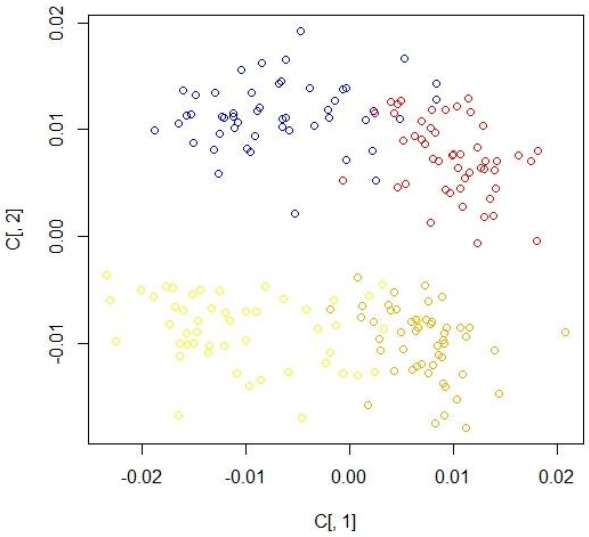
2.3.2 Trouver une solution pour améliorer la qualité de votre représentation en termes de visualisation des différents groupes.

Nous pouvons représenter les 200 crabes en fonction des 5 composantes principales. Cette représentation ne nous permet pas vraiment de distinguer les crabes de différentes espèces et différents sexes. Nous remarquons toutefois qu'il est possible de distinguer les sexes peu précisément (rouge + orange = F, bleu + jaune = M) dans le premier plan factoriel.



Ce premier ACP ne nous permet pas de visualiser des classes distinctes en fonction du sexe et de l'espèce. Nous avons remarqué que toutes les variables dépendaient fortement du premier axe factoriel. Elles sont donc très corrélées. Comme précédemment noté ceci est du à l'effet taille. Afin de mieux distinguer les différents groupes de crabes, nous allons donc normaliser notre jeu de données en divisant les données de chaque individu par la somme de ses données. Nous refaisons ensuite une ACP sur ce nouveau jeu de données. Nous obtenons alors les graphs ci-dessous.

Le premier axe vectoriel explique maintenant 47 % de l'inertie globale et le premier plan factoriel explique 92% de l'inertie globale.
Les variables sont de meilleurs qualités car plus proche du cercle.
On remarque que l'on distingue beaucoup mieux les différents groupes de crabes en fonction de leur espèce et leur sexe (bleu = Male Orange, Rouge = Femelle Orange, Orange = Femelle Blanche, Jaune = Male Blanc).



3 CONCLUSION

Ce TP nous a permis de comprendre plus en détails le fonctionnement et l'utilité de l'ACP. Nous avons pu noter que le logiciel R permet de réaliser une ACP directement à partir d'un jeu de données. Cependant afin d'avoir une représentation des informations pertinentes, il faut être attentif et analyser les données correctement pour ensuite les traiter si besoin. Lors de l'ACP sur les données des crabes, nous avons pu remarquer qu'une ACP effectuée sur les données brutes n'était pas pertinentes dû à l'effet « taille » biaisant les données.