SY09 Printemps 2017
Compte-rendu TP1 : Statistique descriptive, Analyse en composantes principales

Jiawei ZHU - Laëtitia BOUDEREAUX

GI04

5 avril 2017

1 Statistique descriptive

1.1 Notes

1.1.1

Le jeu de données contenu dans le fichier sy02-p2016.csv contient des informations relatives aux étudiants inscrits à lUV SY02 au semestre de printemps 2016. Dans le fichier, nous pouvons compter 296 données sur les étudiants. Chaque donnée décrit 12 attributs : nom de l'étudiant, sa spécialité (branche), son niveau (semestre d'étude), son statut, le dernier diplôme qu'il a obtenu, sa note de médian, le correcteur associé à sa note de médian, sa note de final, le correcteur associé à sa note de final, sa note total et son résultat.

Certaines données dans ce jeu peuvent être manquantes. Trois types de données manquantes existent :
- Dernier diplome obtenu dans le cas des étudiants étrangers
- Notes du médian : si l'étudiant n'est pas venu au médian, il ne peut pas être noté
- Notes du final : même cas que pour la note du médian. Cependant, on remarquera que les personnes qui ne sont pas venus au final ont eu une mauvaise note au médian.

De plus, nous pouvons remarqué que certaines données sont liées : à savoir la note du médian et la note du final vont être à l'origine de la note totale et cette dernière va elle être à l'origine du résultat (A, B, C, D, E ou F).

1.1.2

Grâce à différentes représentations graphiques, nous avons pu étudier plus en profondeur les liens statistiques entre les différentes variables.

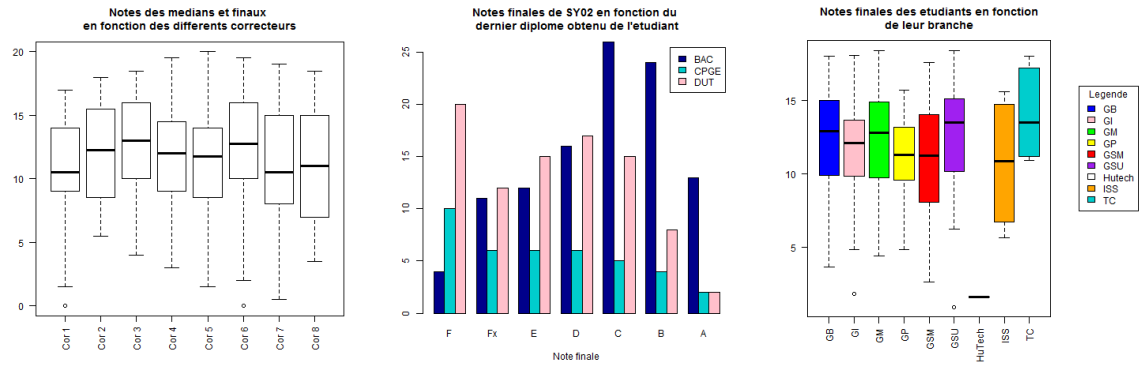

FIGURE 1 – Analyse graphique des liens statistiques entre les différentes variables

1.2 Données crabs

1.2.1

Le jeu de données se compose de 200 observations sur des crabes qui décrivent 7 variables différentes : 2 variables qualitatives et 3 variables quantitatives. Les variables qualitatives décrivent les caractéristiques suivantes : le sexe du crabe (F ou M) et l'espèce du crabe (O ou B). Ci-dessous la représentation des différentes variables qualitatives en fonction des crabes male/female et des crabes d'espèces orange/blue.

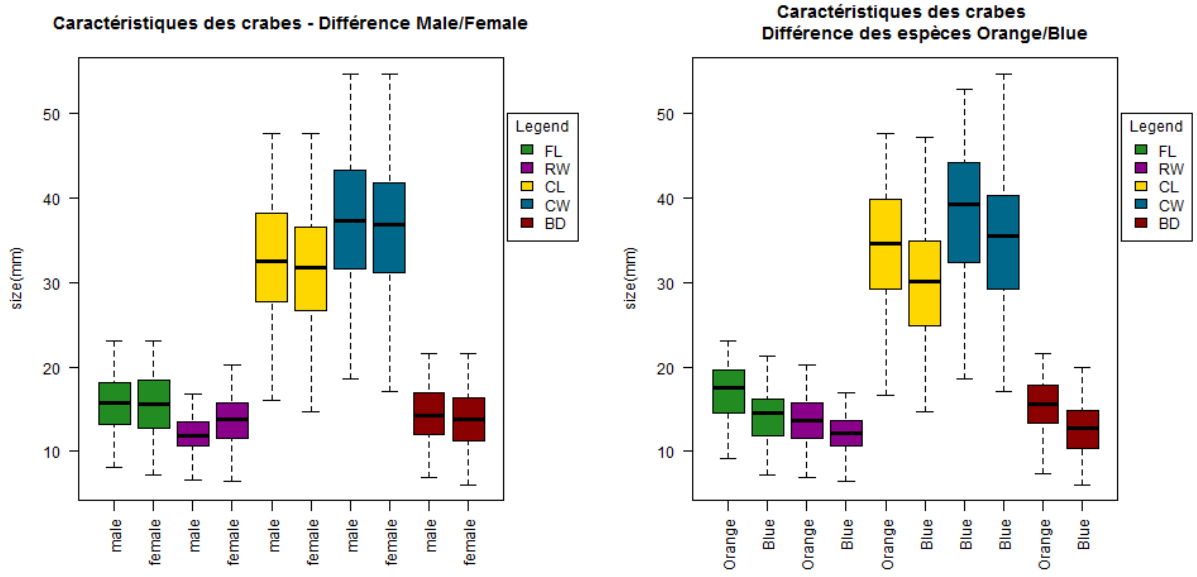

FIGURE 2 – Caractéristiques des crabes

Ce que l'on peut retenir de ces graphiques c'est qu'il n'y a pas une caractéristique qui détermine si une espèce est un mâle ou une femelle et de même pour les espèces. On peut remarquer que quelques caractéristiques sont plus déterminantes que d'autres : par exemple la taille du lobe frontal (FL) est bien plus élevée pour les crabes de l'espèce orange. Cependant pour chaque caractéristique on remarque que les deux boîtes à moustache se chevauchent donc on ne peut pas se fier à cette première analyse pour identifier le sexe ou l'espèce grâce à une des caractéristiques.

Pour aller plus loin dans l'analyse, nous pouvons comparer deux à deux chaque caractéristique grâce à la fonction *plot* sur la matrice *crabsquant*.

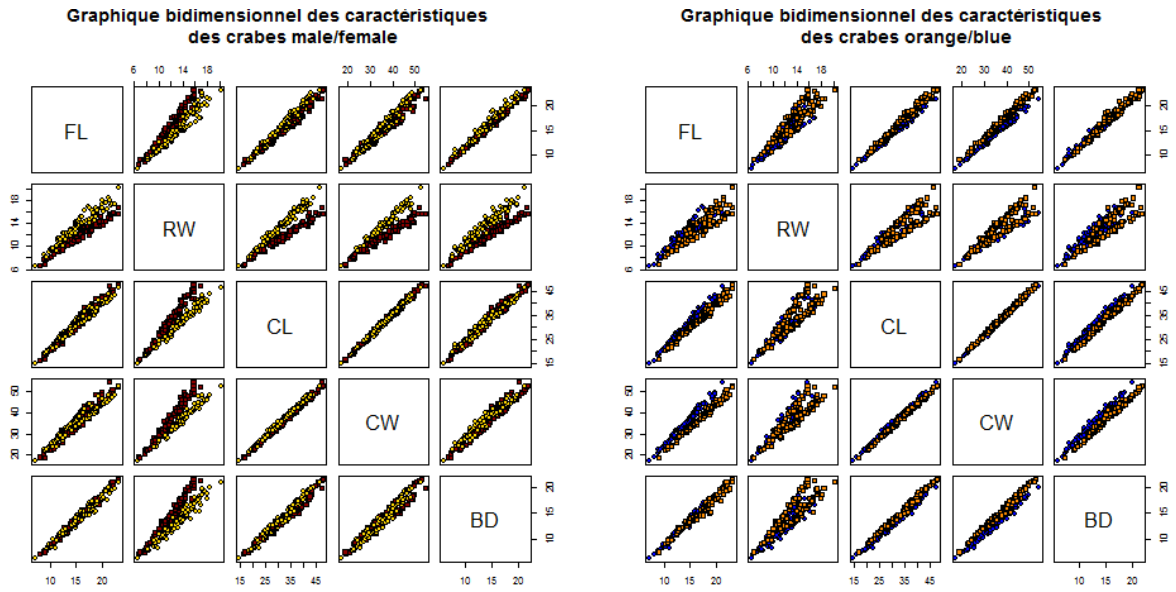

FIGURE 3 – Analyse bidimensionelle

On remarque qu'il existe une forte liaison entre toutes les caractéristiques. Même avec ce graphe bidimensionnel, il est difficile de distinguer les mâles des femelles ou les espèces Orange des espèces Blue.

1.2.2

Pour obtenir la matrice de corrélation du jeu de données *crabsquant*, il suffit d'utiliser la commande *cor*.

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

Nous pouvons remarquer que la matrice de corrélation confirme la forte liaison entre les différentes caractéristiques que nous observions grâce à la figure **??**. En effet, la valeur la plus faible de cette matrice est de 0.889, ce qui est très élevé. On peut expliquer la forte corrélation entre les différentes variables par le fait que la mesure des membres de chaque crabe est proportionnelle : si un crabe est naturellement *grand*, il aura une longue et large carapace mais aussi un grand lobe frontal, ect.

Pour palier à ce problèmes, il suffit d'utiliser l'Analyse en Composantes Principales qui va permettre de transformer des variables liées entre elles en nouvelle variables dé-corrélées les unes des autres que l'on va appeler les composantes principales.

2 Analyse en composantes principales

2.1 Exercice théorique

2.1.1

Pour calculer les axes factoriels de l'ACP du nuage de points défini par ces variables, il faut tout d'abord centrer la matrice *corr.acp*, à savoir soustraire chaque élément des colonnes de variables par la moyenne de cette variable. Pour ceci, nous pouvons utiliser la fonction *scale* et ainsi obtenir la matrice centrée suivante :

$$\mathrm{X} = \begin{pmatrix} -0.01160283 & -0.2760523 & -1.1624801 & 0.1972050 \\ -0.95187094 & -1.7977089 & 1.2278883 & -0.5003495 \\ 0.52689902 & 0.6339053 & -0.3099332 & -1.5786075 \\ 1.56001228 & 0.4388925 & 1.2088317 & 1.0503081 \\ -1.17469929 & -0.0456920 & -0.2385818 & -0.2047466 \\ 0.05126177 & 1.0466553 & -0.7257248 & 1.0361905 \end{pmatrix}$$

Calculer les axes factoriels de lACP du nuage de points défini par les quatre variables quantitatives. Quels sont les pourcentages dinertie expliquée par chacun de ces axes ?

3 Conclusion