Projet LSTAT2110(A) – Analyse de données

PETIT Romain & VAN GEERSDAELE Arthur, 4827-1700 & 3004-1700, GBIO2M

Contents

1 Introduction

En 2003, Grete Heinz, Louis J. Peterson, Roger W. Johnson, et Carter J. Kerk ont mis en place une étude visant à explorer les relations entre les différentes dimensions du corps humain et d'autre caractéristiques.

Pour 507 individus (247 hommes - 260 femmes), 25 mesures/observations, dont les noms et unités sont détaillés en annexe, ont été faites.

Un exemple de problématique très pertinente serait qu'un squelette (entier ou même partiel) soit retrouvé et qu'une enquête soit ouverte. Dans un tel cas de figure, hormis les informations génétiques, les premières caractéristiques intéressantes sont l' *âge*, le *poid*, la *taille* et le *sexe* de l'individu retrouvé.

Ce rapport va analyser ces données en utlisant les méthodes vues au cours afin de voir, si oui ou non, l'idée d'utiliser certaines valeurs de variables pour en déduire d'autres est pertinent.

L'objectif de cette étude était précisément d'offrir aux étudiants des données solides et pertinentes pour les analyser.

2 Présentation des données, analyse descriptive

Il est important de garder en mémoire le fait que les données utilisées ont été mesurées sur des sujets adultes sains et en bonne forme physique.

```
download.file("https://www.openintro.org/book/statdata/bdims.csv", destfile = "bdims.csv")
data.initial <- read.csv("bdims.csv")
data.quantitative <- subset(data.initial, select = -c(length(data.initial)))
```

2.1 Variable d'intérêt : le sexe (0 : Féminin, 1 : Masculin)

Cette variable d'intérêt est discrète et fera l'objet d'une classification plus tard dans le rapport. Le nombre d'hommes et de femmes étudiés est assez équilibré et élevé (247 H - 260 F).

2.2 Variable d'intérêt : l'âge (en année)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
18	23	27	30.18	36	67

2.3 Variable d'intérêt : le poid (en kg)

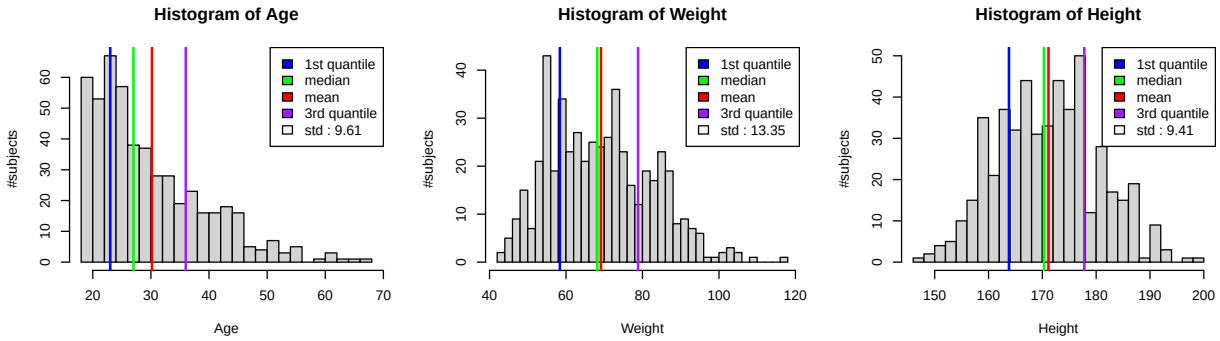
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
42	58.4	68.2	69.15	78.85	116.4

2.4 Variable d'intérêt : la taille (en cm)

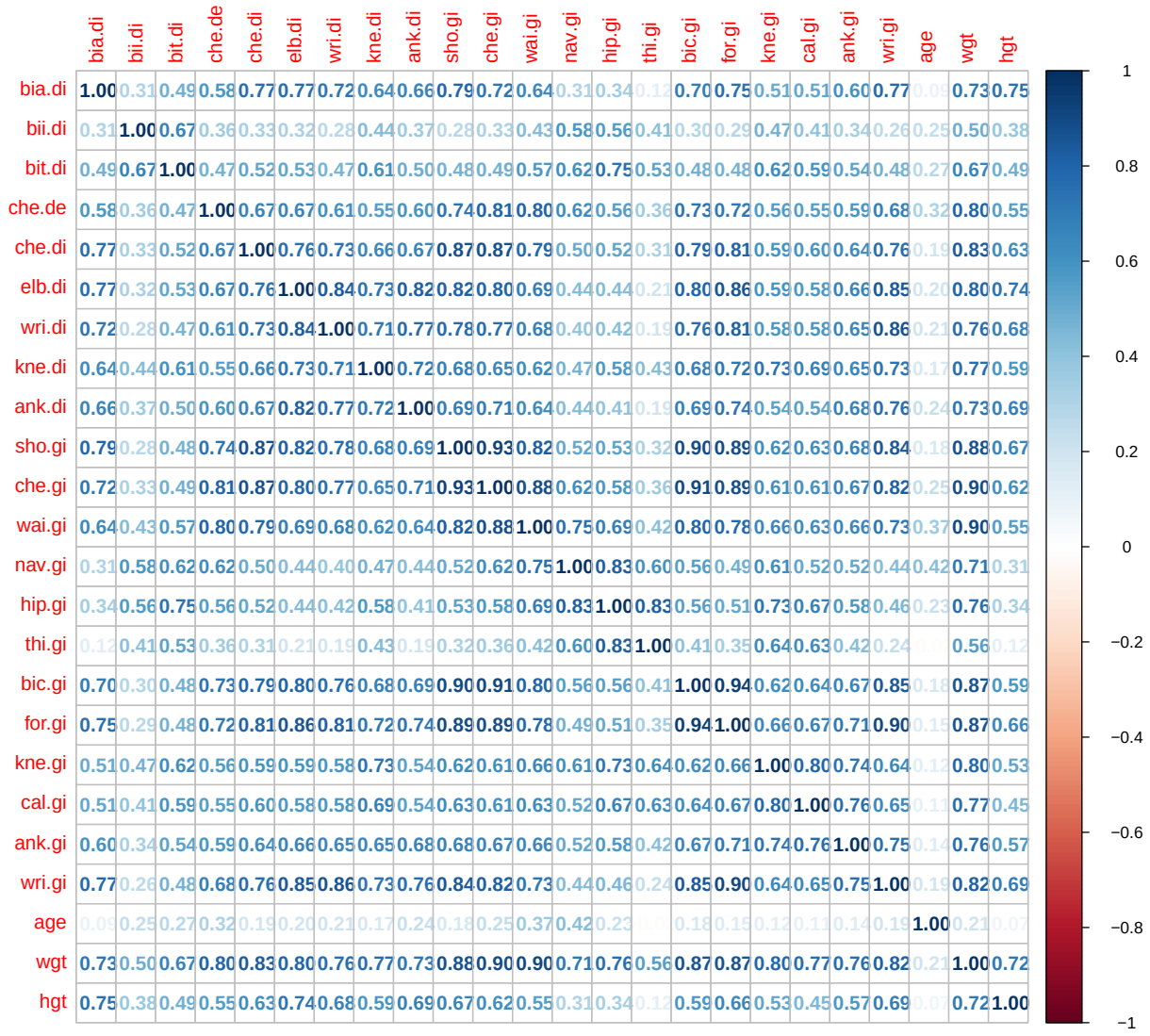
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
147.2	163.8	170.3	171.1	177.8	198.1

2.5 Enlever rigoureusement les individus outliers (à venir, meme si ils semble ne pas vraiment en avoir?)

2.6 Synthèse : distribution des variables d'intérêt

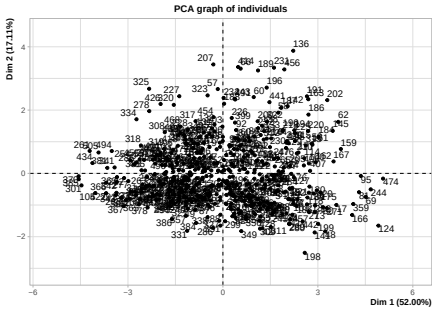


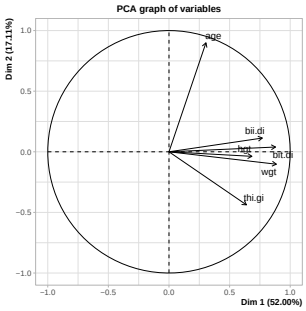
2.7 Correlation entre les variables : corrplot



3 Analyse en composantes principales

Pour simplifer la lecture, les variables hautement correlées sont représentées par une unique flèche.





À première vue de l'ACP, la variable *âge* est fortement décorrelée des autres en regard des autres variables, elle ne saura sans doute jamais être déduite à partir d'autres variables (visible également sur le corrplot). Le graphe des individus de l'ACP montre une population étudiée distribuée homogénéiquement autour de zero, sans outliers sévères.

3.1 Etude de la quantité d'information apportée sur les variables (âge,) poids et taille en fonction des autres variables disponible : (=étude combinée des variances et des correlations)

PCAAAAA (score en variance conservee)

4 Clustering

Faire un cluster avec V1 ... Vn (les variables habituelles) et la "Class" sexe.

inférer sur un individu non-sexualisé

5 Analyse des correspondances

6 Conclusions

7 Annexes

7.1 Définition des données

- bia.di : Un vecteur numérique, le diamètre biacromial du sujet en centimètres.
- bii.di : Un vecteur numérique, le diamètre biiliaque du sujet (largeur pelvienne) en centimètres.
- bit.di : Un vecteur numérique, le diamètre bitrochantérien du sujet en centimètres.

- che.de : un vecteur numérique, la profondeur de la poitrine du sujet en centimètres, mesurée entre la colonne vertébrale et le sternum au niveau du mamelon, à mi-expiration.
- che.di : Un vecteur numérique, le diamètre thoracique du sujet en centimètres, mesuré au niveau du mamelon, à mi-expiration.
- elb.di : Un vecteur numérique, le diamètre du coude du sujet en centimètres, mesuré comme la somme de deux coudes.
- wri.di : Un vecteur numérique, le diamètre du poignet du sujet en centimètres, mesuré comme la somme de deux poignets.

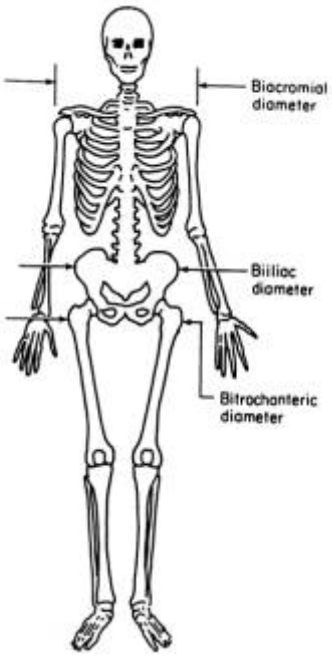


Figure 1: Biacromial, Biiliac, and Bitrochanteric Diameters.

- kne.di : Un vecteur numérique, le diamètre du genou du sujet en centimètres, mesuré comme la somme de deux genoux.
- ank.di : Un vecteur numérique, le diamètre de la cheville du sujet en centimètres, mesuré comme la somme de deux chevilles.
- sho.gi : un vecteur numérique, la circonférence de l'épaule du sujet en centimètres, mesurée sur les muscles deltoïdes.
- che.gi : Un vecteur numérique, le tour de poitrine du sujet en centimètres, mesuré à la ligne du mamelon chez les hommes et juste au-dessus du tissu mammaire chez les femmes, à mi-expiration.
- wai.gi : un vecteur numérique, le tour de taille du sujet en centimètres, mesuré à la partie la plus étroite du torse sous la cage thoracique comme moyenne de la position contractée et détendue.
- nav.gi : un vecteur numérique, la circonférence du nombril (abdominale) du sujet en centimètres, mesurée au niveau de l'ombilic et de la crête iliaque en utilisant la crête iliaque comme point de repère.
- hip.gi : Un vecteur numérique, la circonférence de la hanche du sujet en centimètres, mesurée au niveau du diamètre bitrochantérien.
- thi.gi : Un vecteur numérique, la circonférence de la cuisse du sujet en centimètres, mesurée sous le pli fessier comme la moyenne des circonférences droite et gauche.
- bic.gi : Un vecteur numérique, la circonférence du biceps du sujet en centimètres, mesurée lorsqu'elle est fléchie comme la moyenne des circonférences droite et gauche.
- for.gi : Un vecteur numérique, la circonférence de l'avant-bras du sujet en centimètres, mesurée lorsqu'elle est étendue, paume vers le haut comme moyenne des circonférences droite et gauche.
- kne.gi : Un vecteur numérique, le diamètre du genou du sujet en centimètres, mesuré comme la somme de deux genoux.
- cal.gi : Un vecteur numérique, la circonférence maximale du mollet du sujet en centimètres, mesurée comme la moyenne des circonférences droite et gauche.
- ank.gi : un vecteur numérique, la circonférence minimale de la cheville du sujet en centimètres, mesurée comme la moyenne des circonférences droite et gauche.
- wri.gi : un vecteur numérique, la circonférence minimale du poignet du sujet en centimètres, mesurée comme la moyenne des circonférences droite et gauche.
- age : Un vecteur numérique, l'âge du sujet en années.

- wgt : Un vecteur numérique, le poids du sujet en kilogrammes.
- hgt : Un vecteur numérique, la taille du sujet en centimètres.
- sex : Un vecteur catégoriel, 1 si le sujet est un homme, 0 si une femme.