# 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 cadavre en mauvais état 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 pertinente.

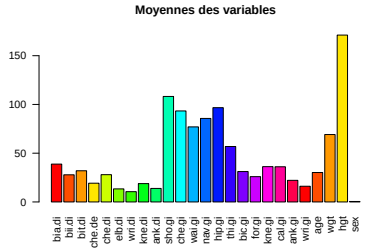
L'objectif de cette étude était précisément d'offrir aux étudiants des données solides et pertinentes pour les analyser.

## 2 Présentation des données, analyse descriptive

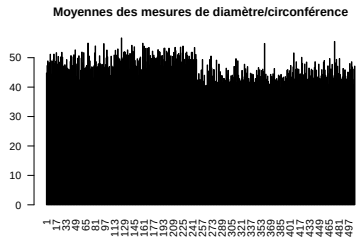
Il est important de garder en mémoire le fait que les données utilisées ont été mesurées sur des sujets adultes sains et en bonne forme physique.

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

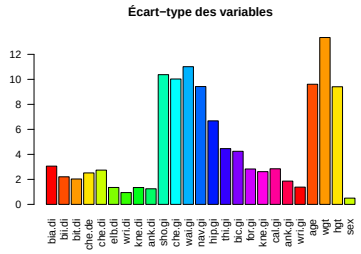
### 2.1 Généralités




On voit sur ce barplot que les moyennes ne sont pas toutes comparables. Le genre (sex) étant une variable binaire (et donc aussi discrète), sa moyenne sera toujours bornée entre zéro et un, alors que pour les autres variables, qui sont des mesures naturelles, ca n'est pas le cas. De plus, les unités de mesure de distance (cm) et de poids (kg) ne sont pas non plus directement comparables. Néanmoins, on remarque déjà de fortes similarités de moyennes entre certains groupe de zones anatomiques (ex: les 9 premières, les 5 suivantes et 7 prochaines).




Ici, on peut observer la moyennes des mesures de diamètre et des mesures de circonférence de chaque individu. Cela ne nous apporte pas grand chose à première vue, mais il suffit déjà de prendre l'information du sexe pour remarquer une différence : les 247 premières moyennes sont des hommes, et les 260 dernières sont des femmes. On peut donc déjà supposer qu'une ou plusieurs de ces mesures sont correlées à la variable catégoricielle du genre.

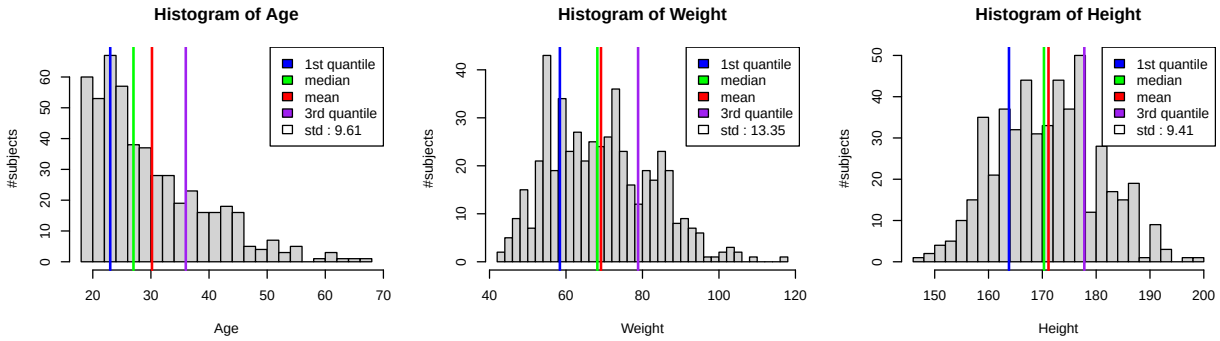



On remarque que les 9 premières mesures anatomiques n'ont que très peu de dispersion autour de leur moyenne, comparé aux 5 suivantes. De nouveau, celles de l'âge, du poids, du sexe et de la taille ne sont pas encore criticable pour les même raisons déjà évoquées ci-dessus.

### 2.2 Variable catégoricielle : *sex* (0 : Féminin, 1 : Masculin)

Cette variable est binaire et discrète. Elle servira à la classification plus tard dans le rapport. Le nombre d'hommes et de femmes étudiés est assez équilibré et élevé (247 H - 260 F).

## 2.3 Distribution des variables de nature différentes (âge, taille et poids)



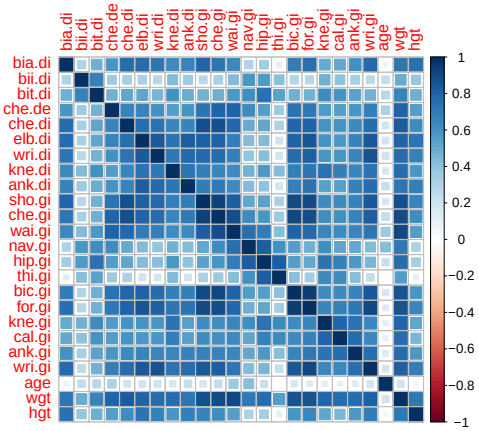

Ici, nous nous intéressons aux variables autres que les mesures anatomiques.

On voit sur le premier graphe que la population est assez jeune (30 ans en moyenne) et que seulement 25% des individus ont plus de 36 ans, tandis que 50% des individus ont moins de 27 ans.

Le second graphe présente presque les mêmes types de quantile et médiane/moyenne que pourrait le faire une Gaussienne. En effet, la moyenne et la médiane sont presque identiques, et les 1er et 2eme quantiles sont équidistants de la moyenne. On note qu'il y'a très peu de personnes au-dessus de 96 kg, dont un sortant vraiment du lot (116 kg).

La distribution de la taille ressemble encore plus à une Gaussienne que celle du poids, avec une moyenne à 1m70 et un écart-type de 9.41 cm, tout genre confondu.

## 2.4 Correlation entre les variables : corrplot

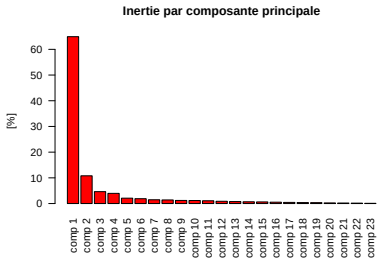


Ce plot des corrélation nous indique malheureusement déjà que nous ne pourrons pas utiliser les mesures anatomiques pour déterminer age (de manière linéaire, car on étudie la corrélation). Étant donné qu'en plus de ça, la distribution de l'âge n'est pas très représentative de la population, nous allons donc l'enlever de la base de données. Si on voulait s'acharner, on pourrait vérifier l'information mutuelle entre la variable âge et toute les autres, afin de détecter une potentielle relation non-linéaire.

Remarque supplémentaire : Au vu de ce plot, les variables : bii.di, thi.gi seront peu intéressantes.

## 3 Analyse en composantes principales

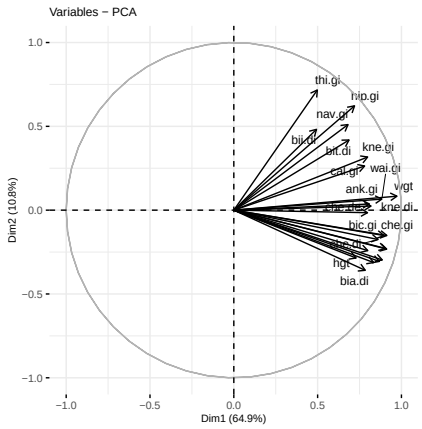
Si on regarde la quantité d'information respectivement apportée par chaque axe dans le tableau ci-dessous, on peut s'en servir pour choisir les composantes principales les plus pertinentes.

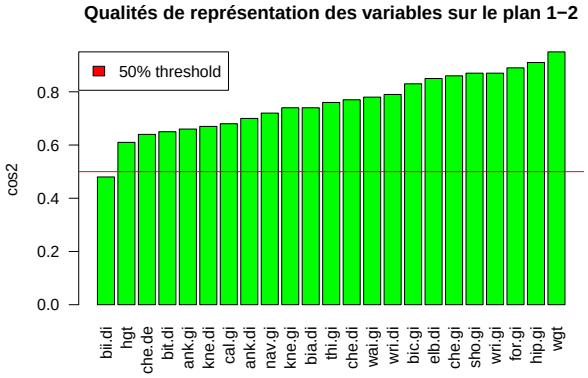



La « règle du coude » nous encourage à ne retenir que les 2 premiers axes qui portent environs 73% de l'information.

La règle de Kaiser nous inciterait à retenir deux axes supplémentaires (pour un gain de 10% d'informations), mais un travail beaucoup plus conséquent.

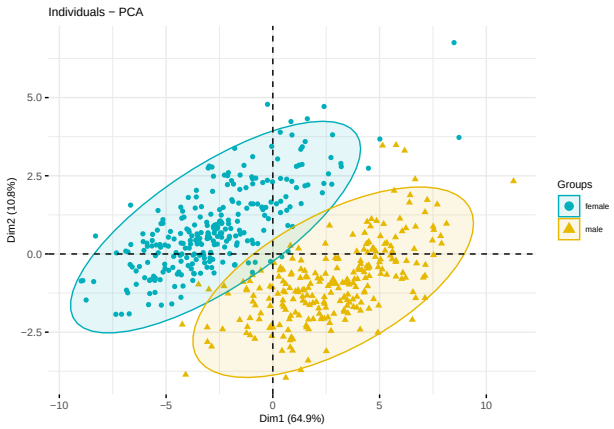
Nous allons donc retenir les deux premières uniquement.




Le cercle des corrélations (et le barplot vert) montre qu'un grand nombre de variables sont plutôt bien représentées car elles sont très proche du cercle, leur qualité de représentation (cos2) est toujours supérieure à 0.5. La seule à être mal représentée est age (cos2 = 0.091), d'ou ses mauvais scores de corrélation avec les autres variables lors du corrplot. Toutes les corrélations sont positives.

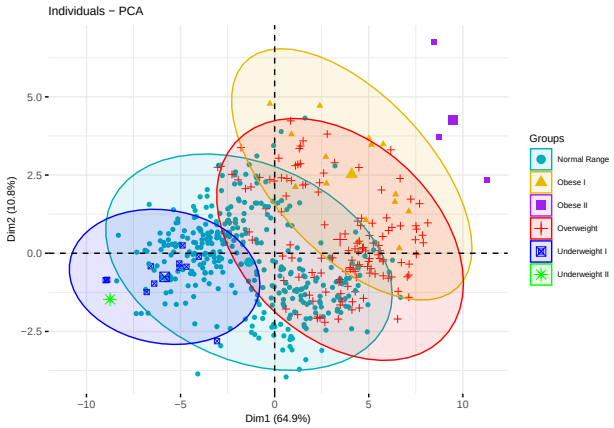
Lors du corrplot, nous avions remarqué que les scores de corrélation bii.di et thi.gi étaient également mauvais. C'est confirmé ici, car on voit qu'ils sont presque perpendiculaires à la majorité des autres variables, plus proches de l'axe 1.




En mappant la variable catégoritielle sex sur le graphe des individus, on peut voir qu'il sont distinctement séparés (hormis quelques outliers) sur les axes 1 et 2. L'information du sexe se trouve donc aussi dans d'autre variables, et le fait de n'avoir gardé que 73% de l'information en sélectionnant que les deux premières composantes ne nous a pas fait perdre cette information.

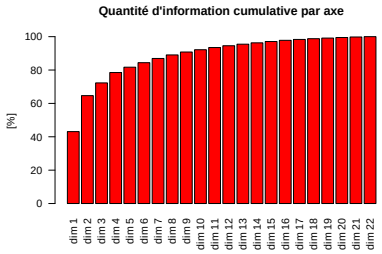
# 4 Clustering

```
Too few points to calculate an ellipse
Too few points to calculate an ellipse
```



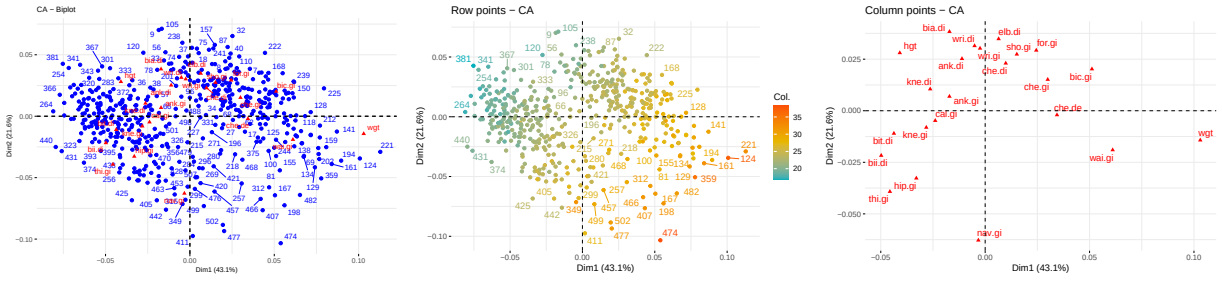
# 5 Analyse des correspondances

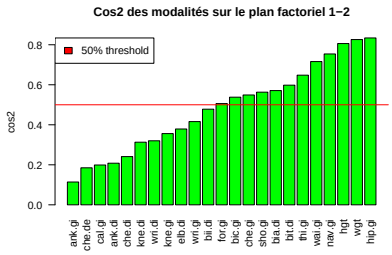
Comme vu en TP, grâce aux relations quasi-barycentriques, nous pouvons représenter sur le même graphique à la fois les coordonnées factorielles des lignes et des colonnes et leur proximité/éloignement a donc une interprétation intrinsèque en termes de liens. Afin de faciliter cette interprétation, on représente généralement les coordonnées de l'ACFS sur le premier plan factoriel. Le graphe de la quantité d'information relative à chaque axe principal ci-dessous nous indique que c'est acceptable, même si prendre également le 3e axe est tentant.




Si maintenant, on affiche le graphe des individus et des variables, on ne sait pas interprêter grand chose, car les individus sont trop nombreux, hormis que le mappage du genre est logiquement conservé avec l'AFC.

Néanmoins, si on mappe la valeur du BMI de chaque individu sur chacun des points du graphe des individus, on peut voir une tendance très nette (S-E vers N-O). Cette tendance se justifie par la position des variables wgt et hgt dans le graphe des variables, malgré que la formule du BMI ne soit pas linéaire. La taille (hgt) a une influence négative sur le BMI, alors que le poids (wgt) a une influence positive.






Sur le barplot des Cos2, hip.gi et le poids (wgt) sont les mieux représentés dans le premier plan factoriel. Un peu moins de la moitié des variables sont sous la barre des 50%, elles ne sont pas assez représentées que pour être analysées. La quantité d'individu est trop vaste que pour afficher le graphe équivalent, mais la qualité de représentation minimale chez eux est de : 0.013 (individu n°488 : une femme de 1m62 et 70kg).

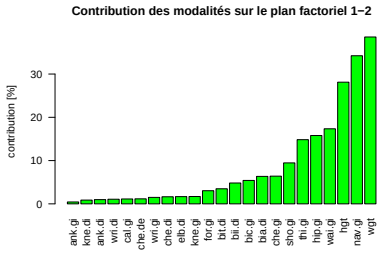



Table 1: Table continues below

| | bia.di | bii.di | bit.di | che.de | che.di | elb.di | wri.di |
|---|---|---|---|---|---|---|---|
| Dim 1 | 0.574 | 3.493 | 3.092 | 1.151 | 0.138 | 0.028 | 0.014 |
| Dim 2 | 5.774 | 1.327 | 0.392 | 0.008 | 1.499 | 1.644 | 1.056 |

Table 2: Table continues below

| | kne.di | ank.di | sho.gi | che.gi | wai.gi | nav.gi | hip.gi |
|---|---|---|---|---|---|---|---|
| Dim 1 | 0.66 | 0.087 | 1.258 | 4.242 | 14.53 | 0.046 | 5.321 |
| Dim 2 | 0.21 | 0.891 | 8.207 | 2.161 | 2.816 | 34.18 | 10.47 |

Table 3: Table continues below

| | thi.gi | bic.gi | for.gi | kne.gi | cal.gi | ank.gi | wri.gi |
|---|---|---|---|---|---|---|---|
| Dim 1 | 5.991 | 4.125 | 0.786 | 1.458 | 1.045 | 0.328 | 0.005 |
| Dim 2 | 8.841 | 1.288 | 2.251 | 0.241 | 0.084 | 0.107 | 1.49 |

| | wgt | hgt |
|---|---|---|
| Dim 1 | 37.14 | 14.48 |
| Dim 2 | 1.425 | 13.63 |

Les variables qui contribuent le plus au premier plan factoriel sont les plus importantes pour expliquer la variabilité du set de données. Les variables qui ne contribuent pas beaucoup à une dimension ou qui contribuent aux dernières dimensions sont moins importantes. Sur le premier axe, la différenciation des individus se fait surtout par le poids wgt, la taille hgt et wai.gi, tandis que sur le deuxième axe, elle se fais principalement par la variable nav.gi.

# 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.

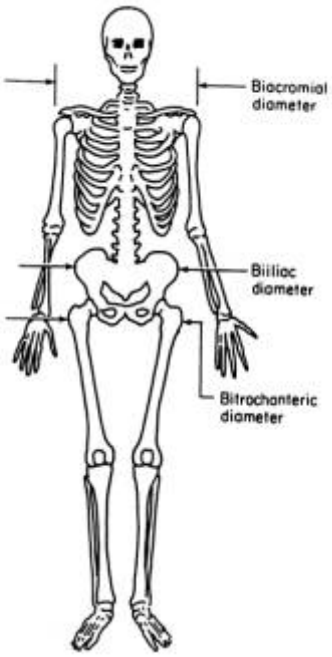



Figure 1: Biacromial, Biiliac, and Bitrochanteric Diameters.

- 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.
- 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.