Formation CNRS : Initiation à R

Une trame d'analyse du « Goldman Data Set »

Frédéric Santos*

21 février 2025

Description des données

On se propose ici de travailler sur un jeu de données anthropométrique disponible en libre accès sur Internet : le « Goldman Data Set ». Ce jeu de données est disponible sur le site de Benjamin Auerbach (`http://web.utk.edu/~auerbach/GOLD.htm`). Il inclut des mesures osseuses de 1538 individus couvrant des périodes allant de -6000 ans jusqu'à nos jours, et provenant de tous les continents (Auerbach, n.d. ; Auerbach & Ruff, 2004). Le détail complet des variables du jeu de données est disponible sur la page web.

1. Inspection, importation et contrôle des données

(i) Télécharger le fichier de données dans le format qui vous semble approprié, puis l'inspecter à l'aide d'un tableur ou d'un éditeur de texte pour en connaître les détails de mise en forme (séparateur décimal, séparateur de champ, indicateur de données manquantes, etc.).

(ii) Importer le fichier avec R en utilisant la fonction adaptée au format de fichier que vous avez choisi. Remarques :

- vous risquez d'éprouver des difficultés pour spécifier le nom des individus, car aucune colonne du jeu de données n'est assimilable à un identifiant unique pour les individus : mieux vaut donc ne pas en spécifier tout de suite ;
- l'encodage du fichier est un encodage `macintosh`. Parcourir l'aide pour savoir comment l'indiquer à R.

(iii) Afficher quelques résumés et effectuer quelques contrôles post-importation [1].

2. Filtrage et mise en forme des données

(iv) Le fichier comporte de très nombreuses variables et nous ne les utiliserons pas toutes. À l'aide de la syntaxe R de base ou d'une fonction du package `dplyr`, retenir uniquement les variables suivantes du data frame : `ID`, `Sex`, `NOTE`, `LHML`, `LHHD`, `LHAPD`, `LHMLD`, `LFML`, `LFHD`, `LRMLD`, `LRAPD`.

*frederic.santos@u-bordeaux.fr

1. Il y en a au moins un qui est assez trivial ici : vous avez l'avantage de connaître le nombre de lignes que le fichier est supposé avoir.

(v) De même, le fichier comporte beaucoup de populations différentes, et nous allons nous concentrer uniquement sur trois populations précises : la population de `Cliff Dweller` (USA, 800–600 BP), la population autrichienne médiévale de `Hainburg`, et la population ancienne de chasseurs-cueilleurs de `Indian Knoll` (USA, 5500–3700 BP). Filtrer le jeu de données pour ne retenir que ces trois populations (indiquées par la colonne `NOTE`). Combien d'individus reste-t-il désormais ?

(vi) La variable `Sexe` est codée de façon peu lisible en l'état. Recoder ce facteur pour transformer les `0` en `Male`; les `1` en `Female`; et transformer les individus dont le sexe est incertain (`0?` ou `1?`) en valeurs manquantes `NA`. Il y a plusieurs possibilités pour cela (cf. l'aide et/ou votre moteur de recherche favori), dont :

- la fonction `levels()`, fonction de base de R;
- la fonction `fct_recode()` du package `forcats`.

(vii) Appliquer un nouveau `summary()` : dans le facteur `NOTE`, certains niveaux de facteurs désormais inutilisés sont toujours présents en mémoire. Bien que ce ne soit pas toujours dérangeant, la fonction `droplevels()` permet de les éliminer (consulter son aide).

3. Analyse de la structure de corrélation

(viii) Calculer et afficher la matrice de corrélations pour l'ensemble des variables numériques du data frame [2]. Détectez-vous « à l'oeil nu » quelque chose de particulier dans la structure de corrélation des variables ?

(ix) Pour aller plus loin, installer (si nécessaire) et charger le package `corrplot`. Consulter sa vignette [3] : `https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html`. À l'aide des informations récoltées sur cette vignette, tenter de mettre en évidence une structure de corrélation par blocs dans ces variables numériques; puis la commenter.

(x) Parmi les paires de variables bien corrélées, il y a la paire `(LFML, LHML)`. Ou bien à l'aide d'une fonction issue du package `lattice` (relativement simple), ou bien à l'aide d'une combinaison de fonctions R de base (plus compliqué !), représenter sur une même fenêtre graphique trois nuages de points. Ils représenteront la liaison `(LFML, LHML)` au sein des trois populations étudiées. Sur chaque nuage, l'information `Sexe` sera mise en évidence par la couleur des points représentés. En résumé, on aura donc un nuage par population, et une couleur par sexe sur chaque nuage.

4. Dimorphisme sexuel pour le fémur à Indian Knoll

(xi) Créer un nouveau data frame nommé `ik`, qui sera constitué uniquement des individus complets en provenance de la population d'Indian Knoll. Combien de femmes et d'hommes reste-t-il ?

(xii) Créer dans ce data frame une nouvelle colonne nommée `Ratio`, qui sera égale au rapport `LFML / LFHD`.

2. D'une manière ou d'une autre, il vous faudra au préalable trouver un moyen de ne retenir que les variables numériques du dataframe.

3. La *vignette* d'un package est un descriptif pédagogique et illustré des fonctionnalités de ce package. De nombreux packages populaires ou « bien maintenus » en possèdent désormais une.

(xiii) Est-ce que ce ratio ainsi créé peut être considéré comme suivant une loi normale à la fois pour les femmes et pour les hommes ? Pour répondre à cette question, on pourra combiner des représentations graphiques (indispensables) avec des tests de normalité (plus facultatifs).

(xiv) Représenter des boites de dispersion en parallèle pour ce ratio en fonction du sexe. Semble-t-il y avoir des différences ? Procéder à un test adéquat pour mettre en évidence ces différences, et commenter également l'intervalle de confiance de la différence entre les deux moyennes.

5. Questions bonus diverses

(xv) Qui est l'individu ayant la valeur maximale pour la variable `LHML` ? Il y aura au moins trois démarches possibles pour répondre à cette question.

(xvi) Il sera souvent utile de définir des fonctions personnelles en R, car tous vos besoins ne seront pas forcément déjà implémentés dans le logiciel ! Voici par exemple comment créer et définir une fonction personnelle qui calcule le coefficient de variation d'un vecteur `x` donné, c'est-à-dire le rapport de son écart-type et de sa moyenne :

```
my_cv <- function(x, na.rm = TRUE) {
    valeur <- sd(x, na.rm = na.rm) / mean(x, na.rm = na.rm)
    return(valeur)
}
```

Utiliser cette fonction pour calculer le coefficient de variation de `LHML`, puis de toutes les variables numériques du jeu de données (la fonction `apply()` aidera beaucoup !).

Références

Auerbach, B. M. (n.d.). *Goldman Osteometric Data Set* [Dr. Auerbach's Personal Website].

Auerbach, B. M., & Ruff, C. B. (2004). Human body mass estimation : A comparison of morphometric and mechanical methods. *American Journal of Physical Anthropology*, *125*(4), 331–342. https://doi.org/10.1002/ajpa.20032