Fiche TD avec le logiciel R : tdr201

Pour une introduction à la statistique descriptive Quelques manipulations dans R

A.B. Dufour & M. Royer

Cette fiche comprend des exercices intégrant à la fois une première approche de la statistique descriptive et la manipulation d'objets dans R.

Table des matières

1 Introduction

Cette fiche de TD, de type `'tdr20.'`, est la première d'un cours d'introduction à la statistique **descriptive** comprenant :

1. la présentation des différents types de variables
2. l'étude d'une seule variable dite analyse univariée
3. les croisements de deux variables dites analyses bivariées.

L'utilisation d'un logiciel permet de raccourcir à la fois le temps de calcul et le temps de réalisation des représentations graphiques.
Le site `http://pbil.univ-lyon1.fr/R/enseignement.html` comprend un ensemble complet de cours et de travaux dirigés ainsi que les jeux de données servant aux analyses.

Le logiciel R est un logiciel de statistique en libre accès sur internet. Il est constitué d'un ensemble de librairies. Chaque librairie est spécialisée dans un secteur de la statistique. Nous n'utiliserons ici que la librairie de base. Il présente également l'avantage d'être multi plate-forme (linux, mac, windows).
R est installé sur tous les ordinateurs des salles du campus de la Doua. Pour débuter une séance de travail,

1. créez un dossier de travail,

2. ouvrez le logiciel R,
3. cherchez dans le menu la rubrique `changer de répertoire de travail` et allez dans le dossier que vous venez de créer,
4. vérifiez que vous êtes bien dans ce dossier avec l'instruction `getwd()`.

2 Exercices

Exercice 1

L'indice de Quételet, appelé encore 'indice de masse corporelle' (IMC) ou 'body mass index' (BMI) est le rapport du poids (en kg) sur la taille (en cm) au carré :

$$IMC = \frac{poids)}{(taille)^2}$$

Il permet de mesurer la corpulence de l'homme adulte. L'Organisation Mondiale de la Santé (OMS) a défini les critères suivants : maigreur (inférieur à 18.5), normal (de 18.5 à 25), risque de surpoids (de 25 à 30), obésité (supérieur à 30).

1) Calculez votre indice IMC.

```
75/(170^2)
```
```
[1] 0.002595156
```

2) Gardez cette valeur en mémoire sous le nom `monIMC`.

```
monIMC <- 75/(170^2)
monIMC
```
```
[1] 0.002595156
```

3) Comparez votre indice aux normes de l'OMS.

Remarque. *On rappelle que l'indice de Quételet n'a qu'une valeur indicative. Pour déterminer l'existence d'une obésité réelle, il faut faire d'autres mesures destinées à établir exactement la proportion de masse grasse, car c'est l'excès de masse grasse qui représente un facteur de risque.*

Exercice 2

Etant un skieur confirmé, vous souhaitez investir dans de l'équipement.

1) Enregistrez dans la variable `budget` la somme que vous acceptez de consacrer à cet achat.
2) Vous êtes intéressé par une paire de skis qui coûte 200 euros, une paire de chaussures de ski vendue 120 euros et une veste polaire à 59 euros. Pouvez-vous réaliser cet achat ?
3) Le gérant du magasin est commerçant et propose de vous offrir la veste polaire si vous achetez les skis et les chaussures. Pouvez-vous réaliser l'achat ?
4) Mais les soldes commencent demain et le gérant offrira 30 % de réduction sur tout le magasin. Quel est le montant total de la réduction ? Quelle est l'offre la plus intéressante ?
5) Après achat, combien d'argent reste-t-il dans votre budget initial ? Modifiez la valeur de la variable `budget` s'il y a eu achat.

Exercice 3

Cinq étudiants, en L3 de la filière STAPS, ont suivi l'UE de Statistique l'année dernière. Les modalités d'examen imposaient un contrôle continu qui comptait pour 40 % de la note et l'examen, qui comptait donc pour 60 % de la note finale.

1) Pour créer le vecteur `ccont` des notes du contrôle continu, tapez la commande suivante :
```
ccont <- c(12, 14, 8, 10, 15.5)
```
2) Les notes de l'examen sont les suivantes : 10, 11, 6, 8.5 et 16. Rentrez ces notes dans le vecteur `exam`.
3) Calculez la note finale des 5 étudiants.
4) Calculez, en utilisant la fonction `mean(x)`, les moyennes du contrôle continu, de l'examen et de l'UE.
5) Tapez la fonction `ls()`. Qu'observez-vous ?

Remarque. Lorsqu'une fonction contient `x`, cela signifie que le nom de l'objet créé, ici le vecteur, doit être rentré dans la parenthèse de la fonction. Par exemple, la moyenne du contrôle continu s'obtient avec la commande `mean(ccont)`.

Exercice 4

Dans l'exercice 1, nous avons calculé l'indice de mase corporelle d'un étudiant.

1) Réaffichez cette valeur.
2) Calculez l'IMC de votre voisin.
3) Rentrez dans le vecteur `taille` la taille des étudiants de votre rangée.
4) Rentrez dans le vecteur `poids` le poids des étudiants de votre rangée.
5) Calculez l'IMC de tous ces étudiants.

Exercice 5

Rentrez dans le vecteur `marathon` les temps en minutes de 15 coureurs :
216, 220, 176, 183, 195, 195, 230, 229, 185, 179, 215, 175, 200, 273, 153

1) Quel est le meilleur temps ? Le moins bon ? [fonctions `max(x)` et `min(x)`]
2) Exprimez le temps de chaque coureur en heures.
3) Sachant que le parcours d'un marathon compte 42.16 km, calculez la vitesse des concurrents.
4) Quelle est la vitesse moyenne ?
5) Tapez la commande `summary(marathon)`. Commentez chacun des résultats obtenus.

Exercice 6

L'intérêt du logiciel R est de manipuler de grands tableaux de données.

1) Pour accéder aux données d'une enquête effectuée sur 237 étudiants, stockées dans le tableau `survey` de la librairie `MASS`, tapez les commandes suivantes :

```
library(MASS)
data(survey)
```

a) Affichez toutes les données de cette enquête en tapant la commande `survey`.

b) La fonction `class(x)` permet de connaitre la nature de l'objet sur lequel nous travaillons. Donnez la nature de l'objet `survey`.

c) Nous pouvons choisir de n'afficher qu'une partie des données comme par exemple les six premières lignes [fonction `head(x)`], les six dernières lignes [fonction `tail(x)`]. Appliquez ces deux fonctions au data frame `survey`.

d) La fonction `names(x)` permet d'afficher le nom des variables du data frame. Affichez le nom des variables de `survey`. Vous trouverez des détails sur les variables étudiées avec la commande `?survey`.

e) Créez un vecteur avec les noms des variables en français : `c( "sexe", "empanME", "empanAutre", "ME", "BrasDessus", "RythmeCard", "Mtape", "exercice", "tabac", "taille", "unites", "age")`. Utiliser ce vecteur pour renommer les variables du tableau [fonction `colnames(x)`].

f) Classer les variables étudiées selon la nature de leurs objets. Notez que chaque variable est **attachée** au tableau étudié. Ce lien se caractérise sous R par le caractère `$`. Ainsi, si nous voulons étudier la variable 'main d'écriture', notée `ME`, nous devons écrire `survey$ME`.

2) Nous allons étudier les variables quantitatives appelées `numeric` sous R.

a) Reprenez la fonction `summary(x)` et étudier chacune de ces variables.

b) Les deux représentations graphiques classiques de l'étude d'une variable quantitative sont l'histogramme [`hist(x)`] et la boite à moustaches [`boxplot(x)`]. Appliquez ces deux fonctions aux variables quantitatives de `survey`. Commentez.

3) Nous allons étudier les variables qualitatives appelées `factor` sous R.

a) Utilisez la fonction `summary(x)` pour étudier chacune de ces variables.

b) Les deux représentations graphiques classiques de l'étude d'une variable qualitative sont la représentation en secteurs ou camembert [`pie(summary(x))`] et la représentation en bâtons [`barplot(summary(x))`]. Appliquez ces deux fonctions aux variables qualitatives de `survey`. Commentez.