Fiche TD avec le logiciel R : tdr202

Introduction à la statistique univariée. Variables et Descriptions générales

A.B. Dufour & M. Royer

Cette fiche comprend des exercices portant sur les paramètres descriptifs principaux des variables quantitatives et qualitatives. Elle s'adresse à des débutants.

Table des matières

1 Introduction

L'objectif est d'acquérir les notions de base concernant la reconnaissance des types de variables (qualitatives et quantitatives) puis de définir les paramètres de position descriptifs qui leur sont associés.

Une liste de documents contenant une information complète pour cette fiche de TD est accessible sur le site `http://pbil.univ-lyon1.fr/R/enseignement.html`,

1. dans le menu `Fiches de TD`, le sous-menu `Le logiciel R`, `tdr13` : Objets.
2. dans le menu `Accueil`, le sous-menu `Page de liens`, `R pour les débutants`
3. dans le menu `Cours`, le sous-menu `Introductions`, de `LANG01` à `LANG04`

2 Variables et Descriptions Générales

Construire un cours de statistique par l'utilisation d'un logiciel comme R, c'est bien sûr s'abstenir du temps de calcul, faciliter la réalisation de graphiques. Mais cela introduit une complexité : la connaissance du vocabulaire et du sens des concepts liés à la statistique d'une part, liés au logiciel d'autre part.

Quelques éléments.

tableau	data frame
variable qualitative	factor
modalité	level
variable quantitative	numeric, integer

2.1 Variable quantitative

Une **série statistique** associée à une variable quantitative X est un liste de valeurs mesurées sur n individus. A chaque individu i est associé la valeur x_i.

Exemple.

```
note <- c(15, 8, 12, 5, 18, 13, 7, 14, 10, 11)
note
 [1] 15  8 12  5 18 13  7 14 10 11
class(note)
[1] "numeric"
length(note)
[1] 10
```

Une **série statistique ordonnée** est une série statistique où les valeurs ont été classées de la plus petite à la plus grande valeur (ou vice-versa). Lorsque le sens de la classification n'est pas précisé, il s'agit de l'ordre croissant. Chaque valeur de la série est notée $x_{(i)}$.

```
sort(note, decreasing = FALSE)
 [1]  5  7  8 10 11 12 13 14 15 18
sort(note, decreasing = TRUE)
 [1] 18 15 14 13 12 11 10  8  7  5
ordnot <- sort(note)
ordnot
 [1]  5  7  8 10 11 12 13 14 15 18
```

Les paramètres classiques associés à une variable quantitative sont :

- le minimum [`min(x)`]
  ```
  min(note)
  [1] 5
  ```
- le maximum [`max(x)`]
  ```
  max(note)
  [1] 18
  ```
- la moyenne : $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ [`mean(x)`]
  ```
  mean(note)
  [1] 11.3
  ```
- Lorsque la série est ordonnée, on peut rechercher les paramètres qui coupent la distribution en plusieurs parties égales. Le cas le plus courant est de couper la distribution en 4 parties : on recherche alors les trois valeurs appelées **quartiles** telles que, à l'intérieur de chaque partie, on retrouve 25% des individus [`quantile(x)`].

```
quantile(note)
  0%   25%   50%   75%  100%
5.00  8.50 11.50 13.75 18.00
```

Ces résultats s'obtiennent également en utilisant une seule fonction de R [`summary(x)`].

```
summary(note)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.00    8.50   11.50   11.30   13.75   18.00
```

Exercice. Calculez les paramètres descriptifs de position sur la série statistique ordonnée. Que constatez-vous ?

La procédure de calcul des quantiles est définie ainsi.

- On note q une des trois valeurs suivantes 0.25, 0.5, 0.75.
- On cherche la position i du quartile dans la série ordonnée : $i = q(n-1)+1$ où n est le nombre total d'individus.
- On repère les deux valeurs de la distribution qui encadrent le quartile cherché.
- On calcule ce dernier par une règle de trois.

Exemple du calcul du premier quartile

```
val <- 0.25
ind <- val * (length(ordnot) - 1) + 1
ind
[1] 3.25
floor(ind)
[1] 3
ordnot[floor(ind)]
[1] 8
ordnot[floor(ind) + 1]
[1] 10
(ordnot[floor(ind) + 1] - ordnot[floor(ind)]) * (ind - floor(ind)) +
    ordnot[floor(ind)]
[1] 8.5
```

Exercice. Calculez, par ce même procédé, les deuxième et troisième quartiles.

2.2 Variable qualitative

Une **série statistique** associée à une variable qualitative A est une liste de valeurs observées sur n individus. Ces valeurs observées sont en nombre restreint et sont appelées les **modalités** de la variable qualitative. A chaque individu i est associée une et une seule modalité. On note p le nombre de modalités de la variable A.

Exemple.

```
mention <- c("Bien", "Echec", "Assez.Bien", "Echec", "Tres.Bien",
    "Assez.Bien", "Echec", "Bien", "Passable", "Passable")
class(mention)
[1] "character"
mention.fac <- factor(mention)
class(mention.fac)
[1] "factor"
```

```
length(mention.fac)
```
```
[1] 10
```

Les paramètres classiques associés à une variable qualitative sont :

- les fréquences absolues c'est-à-dire le nombre d'individus par modalité notées n_k avec k variant de 1 à p,

```
summary(mention.fac)
```
```
Assez.Bien       Bien      Echec   Passable  Tres.Bien
         2          2          3          2          1
```

- les fréquences relatives $f_k = \frac{n_k}{n}$.

```
summary(mention.fac)/length(mention.fac)
```
```
Assez.Bien       Bien      Echec   Passable  Tres.Bien
       0.2        0.2        0.3        0.2        0.1
```

3 Enquête réalisée au près des étudiants de licence 3

Les données du questionnaire ont été rentrées à l'aide d'un tableur (ici, Excel) et sauvegardées en mode `Texte(séparateur :tabulation)(*.txt)(*.txt)`. Le fichier se trouve sur le site de `pbil` dans le menu `Données` et le sous-menu `dossier de fichiers` sous l'appellation `enqL3APA.txt`. Elles contiennent les réponses pour quatre 'générations' d'étudiants : 2006 (lignes 1 à 58), 2007 (lignes 59 à 76) et 2009 (lignes 77 à 96).

1. Vérifiez, à l'aide de la commande `getwd()` que vous vous trouvez bien dans votre espace de travail. Si ce n'est pas le cas, allez dans le menu `Changer de répertoire courant`.
2. Copiez le fichier dans votre dossier de travail (par le clic droit de la souris)
3. Ouvrez Excel ; ouvrez le fichier `enqL3APA.txt` sous Excel ; fermez le tout
4. Allez sous R.

```
enql3apa <- read.table("enqL3APA.txt")
enql3apa
```

Regardez la première ligne du data frame : elle contient les noms des variables. Modifiez la lecture du data frame.

```
enql3apa <- read.table("enqL3APA.txt", header=TRUE)
enql3apa
```

Notez bien la différence entre les deux instructions.

On affiche les lignes correspondant aux étudiants ayant rempli le questionnaire en 2009 que l'on conserve dans un data frame particulier `enq09`.

```
enq09 <- enql3apa[77:96, ]
```

Une variable est **attachée** à un data frame. Le signe de cet attachement est le dollar `$`. Si vous voulez étudier la variable `sexe`, il vous faut nommer la variable comme suit :

```
enq09$sexe
```

Exercice 1

1) Repérez la ligne correspondant à votre réponse au questionnaire. Etes-vous d'accord avec les données enregistrées ? Avez-vous répondu correctement ?

2) Analysez la variable qualitative `sexe`.

Réponse :

```
feminin masculin
      7       13
```

```
feminin masculin
   0.35     0.65
```

3) Analysez la variable quantitative `taille`.

Réponse :

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  153.0   170.0   175.5   173.8   180.0   190.0
```

Exercice 2

1) Donnez la proportion de gauchers en 2009 ?

```
droite gauche
   0.7    0.3
```

2) Mélanger les données homme / femme lorsqu'on travaille sur des mesures n'a pas beaucoup de sens. Afin de ne tenir compte, par exemple, que la taille des femmes, l'instruction est la suivante :

```
enq09$taille[enq09$sexe == "feminin"]
```

```
[1] 172 153 160 170 176 162 168
```

Calculez, pour les variables taille et poids, les paramètres statistiques élémentaires des hommes et des femmes.

Réponse attendue pour la taille des hommes :

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  170.0   175.0   179.0   178.2   180.0   190.0
```

Réponse attendue pour la taille des femmes :

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  153.0   161.0   168.0   165.9   171.0   176.0
```

Réponse attendue pour le poids des hommes :

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  58.00   65.00   70.00   72.31   80.00   90.00
```

Réponse attendue pour le poids des femmes :

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  48.00   51.00   54.00   55.29   58.00   67.00
```

3) Tapez l'instruction ci-dessous et commentez les résultats obtenus.

```
summary(enq09)
```

```
    annee       groupe  identifiant          sexe         poids           taille     
Min.   :2009   A:20   Min.   : 1.00   feminin : 7   Min.   :48.00   Min.   :153.0  
1st Qu.:2009   B: 0   1st Qu.: 5.75   masculin:13   1st Qu.:58.00   1st Qu.:170.0  
Median :2009          Median :10.50                 Median :65.50   Median :175.5  
Mean   :2009          Mean   :10.50                 Mean   :66.35   Mean   :173.8  
3rd Qu.:2009          3rd Qu.:15.25                 3rd Qu.:73.00   3rd Qu.:180.0  
Max.   :2009          Max.   :20.00                 Max.   :90.00   Max.   :190.0  
  rythmcard           age      baccalaureat mention      hmental       hmoteur     
Min.   :56.00   03/08/88: 1   S      :9     AB  : 5   Min.   :0.0   Min.   :0.00  
```

```
1st Qu.:63.25   03/12/84: 1   ES     :5     B   : 1   1st Qu.:0.0   1st Qu.:0.00
Median :66.50   03/12/89: 1   STL    :2     P   :11   Median :0.0   Median :1.00
Mean   :66.75   04/02/86: 1   SMS    :1     NA's: 3   Mean   :0.4   Mean   :0.65
3rd Qu.:70.50   07/04/89: 1   STT    :1               3rd Qu.:1.0   3rd Qu.:1.00
Max.   :84.00   (Other) :13   (Other):0               Max.   :1.0   Max.   :1.00
                NA's    : 2   NA's   :2
  hsensoriel     pblesocial  pratique         sport          niveau    mecriture
Min.   :0.00   Min.   :0.0   non: 8   escalade   :2   competition: 8   droite:14
1st Qu.:0.00   1st Qu.:0.0   oui:12   football   :2   loisir     :10   gauche: 6
Median :1.00   Median :0.0            athletisme :1   NA's       : 2
Mean   :0.55   Mean   :0.3            basket     :1
3rd Qu.:1.00   3rd Qu.:1.0            boxe_France:1
Max.   :1.00   Max.   :1.0            (Other)    :5
                                      NA's       :8
mfourchette   pballon       oeil     rotation     pappui
droite:12   droit :15   droit :9   droit :10   droit :12
gauche: 7   gauche: 5   gauche:9   gauche:10   gauche: 7
NA's  : 1               NA's  :2               NA's  : 1
```