Chapitre **2** **Le Logiciel R et Statistique descriptive**

# Introduction

Ce chapitre décrit les différentes commandes à taper sous R pour structurer les variables, tracer des graphiques et calculer des résumés numériques statistiques simples sur un jeu de données. Nous allons fonder tous les exemples de ce chapitre sur le fichier de données *imcenfant.csv qui est créé dans le logiciel excel sous l'extension csv(séparateur :point-virgule)*

## 2.1 Représentation Brute et Importation des fichiers

### 2.1.1 Présentation du fichier:

Un échantillon de dossiers d'enfants a été saisi. Ce sont des enfants vus lors d'une visite en France section de maternelle en 1996-1997 dans des écoles de Bordeaux (Gironde, France). L'échantillon est constitué de 152 enfants âgés de 3

Variables et codage

| **Description** | **Unite ou Codage** | **Variable** |
|---|---|---|
| Sexe | F pour fille ; G pour garçon | SEXE |
| Ecole située en zone d'éducation prioritaire | O pour oui; N pour non | zep |
| Poids | Kg (arrondi a 100 g pres) | poids |
| Age à la date de la visite | Années | an |
| Age à la date de la visite | Mois | mois |
| Taille | En cm (arrondi à 0,5 cm pres) | taille |

## 2.2 Structuration des variables suivant leur type

### 2.2.1 Structurer les variables qualitatives

Pour les variables qualitatives, la structure est imposée au moyen de la fonction **as. factor ()**. Il peut éventuellement être intéressant d'utiliser aussi la fonction **levels ()** pour recoder les modalités d'une variable qualitative.

Effectuons ces opérations sur les variables qualitatives de notre jeu de données (cf ; Tp1)

```
> sexe <- as.factor(SEXE)
> levels (sexe) <- c ("Garçons", "Fille")
```

### 2.2.2 Structurer les variables ordinales

Pour les variables ordinales, la structure est imposée au moyen de la fonction **as. ordered()**. Il peut éventuellement être intéressant d'utiliser aussi la fonction **levels ()** pour recoder les modalités d'une variable ordinale.

Effectuons ces opérations sur les variables ordinales de notre jeu de données:

```
> zep <- as.ordered(zep)
> niveaux <- c ("Oui", "Non")
> levels(zep) <- niveaux
```

### 2.2.3 Structurer les variables quantitatives discrètes

Pour une variable discrète, la structure est imposée au moyen de la fonction **as. integer ()** .

```
> an <- as. integer (an)
> moi <- as. integer (mois)
```
Cela n'est toutefois valable que si les données observées sont des entiers.

### 2.2.4 Structurer les variables quantitatives continues
Pour une variable continue, la structure est imposée au moyen de la fonction **as. double ()**
```
> tai <- as.double(taille)
> poi <- as.double(poids)
```

## 2.3 Tableaux de données
### 2.3.1 Tableaux des données individuelles
Il s'agit du type d'organisation le plus courant. On dispose des mesures d'une ou de plusieurs variables pour chacun des *N* individus constitutifs d'une certaine population.

### 2.3.2 Tableaux des effectifs ou des fréquences d'une variable
Il est souvent intéressant de représenter un tableau de données individuelles (ou tableau de données brutes) sous une forme plus condensée. Ainsi, le tableau des effectifs ou des fréquences (appelé aussi tri à plat) permet d'appréhender plus facilement la distribution d'une variable, notamment qualitative ou ordinale. Il s'obtient au moyen de la fonction **table()**.
**EXemple**
```
> tpe <- table(sex)# Tri à plat en effectifs.
> tpe

> tpf <- tpe/length(sex)      # Tri à plat en fréquences.
> tpf
> levels(sex)       # Modalités.
> nlevels(sex)      # Nombre de modalités.
```

### 2.3.3 Tableaux de données regroupées en classes
Il est parfois intéressant de représenter un tableau de données individuelles (ou tableau de données brutes), récoltées sur une ou plusieurs variables quantitatives, sous une forme plus condensée. On utilise pour cela un tableau de données regroupées en classes, en notant les effectifs (ou les fréquences) de différentes classes préalablement déterminées. Pour cela il faut les instructions (l'exemple de taille étant inclus :
```
> res <- hist(tai,plot=FALSE)
> nn <- as.character(res$breaks)
> x <- as.table(res$counts)
> dimnames(x) <- list(paste(nn[-length(nn)],nn[-1],sep="-"))
```

## 2.3 Résumés Numériques
Nous présentons, pour plus de simplicité, tous les résumés numériques sur le vecteur $X = (X_1, \ldots, X_N)^t$. Ce vecteur est l'ensemble des $N$ valeurs de la variable $X$ mesurées sur une population d'effectif $N$ (cas standard de la statistique descriptive). Les exemples d'application seront principalement fondés sur la série de données du vecteur taille.
REMARQUE Les résumés numériques ne peuvent être calculés en présence de données manquantes (NA). Si cela est nécessaire, il est possible d'utiliser la fonction **na. omit ()** pour les retirer lors du calcul.
```
> x <- na.omit(taille) # Inutile ici car taille n'a pas de NA.
```

### 2.3.1 Resumes de position d'une distribution
#### *2.3.1.1 Le ou les modes*
Les modes sont les valeurs de la variable X qui apparaissent le plus fréquemment. Ils peuvent se calculer pour une variable de n'importe quel type, bien que, pour une variable quantitative continue, on présente plutôt la classe modale. Notez que le mode peut être unique, auquel cas on parle de distribution unimodale, par opposition à des variables multimodales.

```
> names(which.max(table(the)))      # Obtention d'un mode unique.
> names(table(the)) [max(table(the))==table(the)]      # Obtention de tous les modes.
```

Ici, la variable the (nombre de tasses de the par jour) est unimodale. Dans le cas d'une variable quantitative, vous pouvez utiliser la fonction **as. numeric ()** sur les résultats ci-dessus pour récupérer des valeurs numériques.

#### *2.3.1.2 La médiane*

La médiane d'une série statistique est la valeur *me* de la variable X qui partage cette série statistique en deux parties (inferieure et supérieure à *me)* de même effectif, les valeurs du caractère étant rangées dans l'ordre croissant. Il s'agit d'un critère de position qui ne se calcule évidemment pas pour des variables purement qualitatives. Pour la calculer, on distingue deux cas:

-l'effectif total *N* de la série est impair. Dans ce cas, la médiane est la valeur située à la position $\frac{N+1}{2}$ ;

-l'effectif total *N* de la série est pair. Dans ce cas, n'importe quelle valeur comprise entre les valeurs aux positions $\frac{N}{2}$ et $\frac{N}{2}$ + 1 peut être considérée comme une médiane de la série. En pratique, la médiane est généralement la moyenne de ces deux valeurs.

La fonction R permettant de calculer une médiane uniquement pour des données numériques est median ().

```
> median (taille)
```

Voici le calcul de la médiane pour les données du vecteur x regroupées en classes au moyen de la fonction **hist ()** :

```
> res <- hist(x,plot=FALSE,breaks=c(130,150,160,170,180,190))
> tab.x <- table(rep(res$breaks[-l],res$counts))
> tab.x
> mediane.sur.freq(tab.x/sum(tab.x))
```

#### *2.3.1.3 La moyenne*

Elle se calcule uniquement pour des variables quantitatives.

```
> mean (x )
```

**Remarque** : La fonction **summary()** appliquée à un vecteur de données quantitatives permet de calculer le minimum, le maximum, la moyenne et les trois quartiles.

### 2.3.2 Paramètres de Dispersion d'une Distribution

Ces paramètres peuvent être calculés uniquement pour des variables quantitatives. Nous les présentons dans le tableau ci-dessous après avoir défini les trois fonctions R suivantes :

```
> var.pop <- function(x) var(x)*(length(x)-l)/length(x)      # Variance de la population
> sd.pop <- function(x) sqrt(var.pop(x))          # Ecart type de la population
> co.var <- function(x) sd.pop(x)/mean(x)        # Coefficient de variation
```

Une estimation sans biais de la variance de la population, fondée sur un échantillon de taille *n,* est calculée au moyen de la fonction var(). L'écart-type correspondant est calculé au moyen de la fonction sd () :

```
> var(x)
> sd(x)
```

## 2.4 Représentations graphiques

Rappelons qu'il convient toujours de choisir adéquatement le mode de représentation graphique d'une variable adapté à son type. En effet, le type d'une variable est souvent traduit par des caractéristiques particulières d'un graphique donné.

### 2.4.1 Graphiques pour les variables qualitatives

**L'**obtention se fait au moyen de la fonction barplot () respectivement la fonction pie(). On obtient le diagramme en barres pour les variables ordinales et le diagramme circulaire pour les variables nominales :

```
> col <- c ("gray", "orangered")
>pie(table(sex),col=col)
> barplot(table(zep),col=col)
```

### 2.4.1 Graphiques pour les variables quantitatives

**L'**obtention se fait au moyen de la fonction plot () respectivement la fonction hist(). On obtient le diagramme en batons pour les variables discrètes et 'histogramme pour les variables continues :

```
>pie(table(an),col="gray")
> hist(table(tai),col="orangered")
```

.