Mémento R

Structure de données

En règle générale, on travaillera à partir d'un tableau rectangulaire, chaque ligne désignant une unité statistique regroupant un ensemble d'observations attachées à des variables arrangées en colonnes, appelé "data frame" (fig. 1).

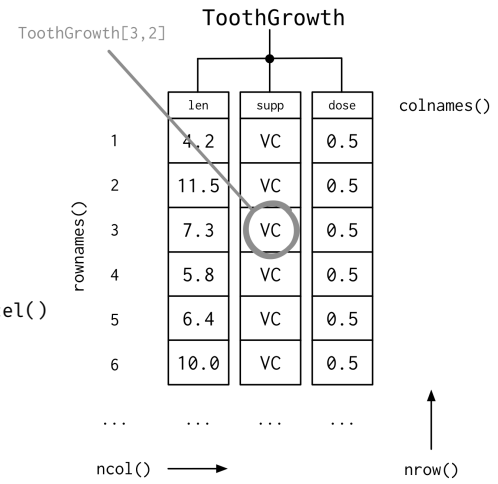

FIG. 1: Structure d'un data frame

Propriétés d'un data frame

```
data(ToothGrowth)            # chargement de données internes
str(ToothGrowth)             # structure d'un objet R
head(ToothGrowth, n = 5)     # en-tête d'un objet R
```

Les fichiers Excel, CSV et texte peuvent être importés à l'aide des commandes `readxl` :: `read_excel()` et `read.table()`. Le package `haven` permet de lire les fichiers SPSS, Stata (incluant v.13+) et SAS.

Importation de fichiers texte

```
setwd("~/Desktop")                # changement de répertoire
dir()                             # liste des fichiers présents
d <- read.table("exemple.csv",    # nom du fichier
                sep = ";",        # séparateur de champ
                dec = ",",        # séparateur décimal
                na.strings = ".", # codage des valeurs manquantes
                stringsAsFactors = FALSE)
```

Les variables catégorielles sont représentés sous forme de facteurs, avec des niveaux (ordonnés ou non) et des étiquettes textuelles associées à chaque niveau.

Codage des variables catégorielles

```
ToothGrowth$dose <- factor(ToothGrowth$dose, labels = c("low", "mid", "high"))
levels(ToothGrowth$dose)[1:2] <- "mid+low"  # aggrégation des deux premiers niveau
as.numeric(ToothGrowth$dose)                # codes numériques (1/2)
```

Représentations graphiques

Le package `ggplot2` repose sur un principe de superposition de couches graphiques et associe : un data frame et un mapping (esthétique) entre différentes variables, un ou plusieurs objets géométriques, un système de facettes, des échelles pour les axes et un système de coordonnées, des annotations pour les axes et un thème graphique (fig. 2).

Graphique d'interaction

```
library(ggplot2)
p <- ggplot(data = ToothGrowth, aes(x = dose, y = len, color = supp)) +
  geom_point(size = 1, position = position_jitter(width = 0.05)) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Dose (mg/day)", y = "Length (oc. units)") +
  theme_bw()
p
```

Modèles statistiques

La commande `lm()` permet d'estimer les paramètres d'un modèle linéaire. Pour un modèle linéaire généralisé (e.g., régression logistique), on utilisera `glm()` en spécifiant l'échelle de lien et la distribution des erreurs : `glm(y ~ x, data = d, family = binomial("logit"))`. Il est commode de stocker les résultats d'un modèle dans une variable afin d'utiliser les commandes associées (fig. 3).

ANOVA et régression linéaire multiple

```
summary(aov(len ~ supp + factor(dose) + supp:factor(dose), data = ToothGrowth))
m0 <- lm(len ~ supp, data = ToothGrowth)
m1 <- lm(len ~ supp + dose, data = ToothGrowth)
anova(m0, m1)
```

Objets géométriques de base

`geom_bar()`, `geom_histogram()`, `geom_point()`, `geom_density()`, `geom_boxplot()`, `geom_smooth()`

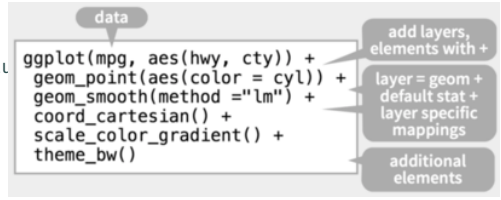

FIG. 2: Exemple de commandes ggplot

Commandes pour les tests d'association usuels

`cor.test()` (test pour un coefficient de corrélation linéaire), `chisq.test()` (test du χ^2 de Pearson), `fisher.test()` (test exact de Fisher), `prop.test()` (test de proportion(s) utilisant une approximation par la loi normale), `t.test()` (test de Student, cas de deux échantillons indépendants ou non (`paired = TRUE`)), `wilcox.test()` (test de Mann-Whitney-Wilcoxon pour deux échantillons indépendants et test des rangs signés de Wilcoxon pour deux échantillons non indépendants (`paired = TRUE`))

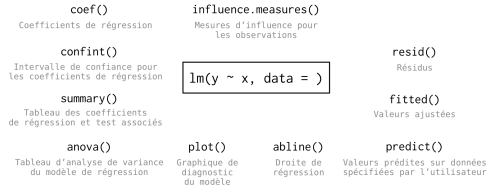

FIG. 3: Commandes associées aux modèles linéaires