Données manquantes : (petite) introduction

Vincent Guillemot
Amaury Vaysse

Avant de commencer

Nous aurons besoin de charger les librairies suivantes :

```
library(dplyr)
#>
#> Attaching package: 'dplyr'
#> The following objects are masked from 'package:stats':
#>
#>     filter, lag
#> The following objects are masked from 'package:base':
#>
#>     intersect, setdiff, setequal, union
library(tidyr)
```

D’où ça vient ?

- Questionnaire : “ne se prononce pas”
- “Erreur” de mesure
- Données perdues
- Opérations interdites

Comment ça se présente ?

R	Description	Exemple
`NaN`	Le résultat impossible (e.g.)	`1 / 0`
`NULL`	L'objet vide	`fruits$umami`
`""`	La chaîne de caractères vide	`""`
`NA`	La vraie donnée manquante	`x <- c(NA, 2, 3)`

Et quel effet cela a ?

Valeur manquante

Opération	Résultat
`3 + NA`	NA
`NA/2`	NA
`TRUE & NA`	NA
`TRUE | NA`	TRUE
`x + 1`	[1] NA 3 4
`sum(x)`	[1] NA

NaN

Opération	Résultat
`3 + NaN`	NaN
`NaN/2`	NaN
`TRUE & NaN`	NA
`TRUE | NaN`	TRUE

Construire son exemple

L'intérêt de construire un petit exemple est de tester des fonctions qui ne nous sont pas familières!

```
fruits_na <- tibble(
  name = c("Apple", "Banana", "Cherry", "Date", "Elderberry", "Fig", "Grape"),
  sugar = c(10.3, 17.2, NA, 63.3, 6.5, 16.2, 16.0),
  # sugar content in g/100g
  water = c(86, 74, 82, 20, 80, NA, 81)
  # water content as a percentage
)
```

J'ai demandé à ChatGPT de créer un petit exemple

Comment on gère ?

Enlever les observations avec données manquantes

```
fruits_na %>% drop_na()
#> # A tibble: 5 × 3
#>   name       sugar water
#>   <chr>      <dbl> <dbl>
#> 1 Apple       10.3    86
#> 2 Banana      17.2    74
#> 3 Date        63.3    20
#> 4 Elderberry   6.5    80
#> 5 Grape       16      81
```

Remplacer les observations avec données manquantes

```
fruits_na %>% replace_na(list(sugar = 0, water = 1))
#> # A tibble: 7 × 3
#>   name       sugar water
#>   <chr>      <dbl> <dbl>
#> 1 Apple       10.3    86
#> 2 Banana      17.2    74
#> 3 Cherry       0      82
#> 4 Date        63.3    20
#> 5 Elderberry   6.5    80
#> 6 Fig         16.2     1
#> 7 Grape       16      81
```

Utiliser des fonctions qui peuvent enlever les valeurs manquantes

```
fruits_na %>% summarize(
  MeanSugar = mean(sugar, na.rm = TRUE),
  MeanWater = mean(water, na.rm = TRUE),
  MedianSugar = median(sugar, na.rm = TRUE),
  MedianWater = median(water, na.rm = TRUE))
#> # A tibble: 1 × 4
#>   MeanSugar MeanWater MedianSugar MedianWater
#>       <dbl>     <dbl>       <dbl>       <dbl>
#> 1      21.6      70.5        16.1        80.5

cor(fruits_na$sugar, fruits_na$water, use = "complete.obs")
#> [1] -0.985215
```

Aller plus loin

- Visualisation des données manquantes avec le package `naniar`
- Imputation de données manquantes avec le package `mice`
- La *Task View* sur les données manquantes