Labo5: Statistiques descriptives univariées

Visseho Adjiwanou, PhD.

06 February 2020

Enquête sociale générale, 1996

- Il s'agit du CROP Socio-Cultural Survey de 1996
- Dans cette partie, nous allons apprendre à :
  - Selectionner les variables
  - Sélectionnez les observations
  - Réorganiser les données
  - Créer de nouvelles variables avec des fonctions de variables existantes (mutate ())
  - Recoder des variables existantes
  - Calculer des statistiques univariées

Dressons la table

```
# Effacer votre environnement

rm(list = ls())

# Installer les package dont vous avez besoin
#install.packages("tidyverse")
#install.packages("summarytools")

# install.packages("tidyverse")
# install.packages("summarytools")

# Charger les packages - Étape fondamentales

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------- tidyverse 1.2.1 --
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## -- Conflicts ------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(summarytools)
```

```
## Registered S3 method overwritten by 'pryr':
##   method      from
##   print.bytes Rcpp
##
## Attaching package: 'summarytools'
```

```
## The following object is masked from 'package:tibble':
##
##     view
```

Téléchargement de la base de données

```
crsc96 <- read_csv("cora-crsc1996-E-1996_F1.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

Regardons ce que contient cette base de données

```
# trois manière de faire
#View(crsc96)

head(crsc96)
```

```
## # A tibble: 6 x 416
##    sexq  ageq commsize region   age    q1    q2    q3    q4    q5    q6
##   <dbl> <dbl>    <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     2     3        1      9    33     1     5     5     5     5     5
## 2     2     3        1      9    34     2     5     4     5     1     5
## 3     2     4        1      9    56     2     2     4     5     5     5
## 4     1     5        1      9    69     1     4     2     4     2     4
## 5     1     3        1      9    43     4     4     4     5     2     5
## 6     2     2        1      9    28     4     5     4     5     1     5
## # ... with 405 more variables: q7 <dbl>, q8 <dbl>, q9 <dbl>, q10 <dbl>,
## #   q11 <dbl>, q12 <dbl>, q13 <dbl>, q14 <dbl>, q15 <dbl>, q16 <dbl>,
## #   q17 <dbl>, q18 <dbl>, q19 <dbl>, q20 <dbl>, q21 <dbl>, q22 <dbl>,
## #   q23 <dbl>, q24 <dbl>, q25 <dbl>, q26 <dbl>, q27 <dbl>, q28 <dbl>,
## #   q29 <dbl>, q30 <dbl>, q31 <dbl>, q32 <dbl>, q33 <dbl>, q34 <dbl>,
## #   q35 <dbl>, q36 <dbl>, q37 <dbl>, q38 <dbl>, q39 <dbl>, q40 <dbl>,
## #   q41 <dbl>, q42 <dbl>, q43 <dbl>, q44 <dbl>, q45 <dbl>, q46 <dbl>,
## #   q47 <dbl>, q48 <dbl>, q49 <dbl>, q50 <dbl>, q51 <dbl>, q52 <dbl>,
## #   q53 <dbl>, q54 <dbl>, q55 <dbl>, q56 <dbl>, q57 <dbl>, q58 <dbl>,
## #   q59 <dbl>, q60 <dbl>, q61 <dbl>, q62 <dbl>, q63 <dbl>, q64 <dbl>,
## #   q65 <dbl>, q66 <dbl>, q67 <dbl>, q68 <dbl>, q69 <dbl>, q70 <dbl>,
## #   q71 <dbl>, q72 <dbl>, q73 <dbl>, q74 <dbl>, q75 <dbl>, q76 <dbl>,
## #   q77 <dbl>, q78 <dbl>, q79 <dbl>, q80 <dbl>, q81 <dbl>, q82 <dbl>,
## #   q83 <dbl>, q84 <dbl>, q85 <dbl>, q86 <dbl>, q87 <dbl>, q88 <dbl>,
## #   q89 <dbl>, q90 <dbl>, q91 <dbl>, q92 <dbl>, q93 <dbl>, q94 <dbl>,
## #   q95 <dbl>, q96 <dbl>, q97 <dbl>, q98 <dbl>, q99 <dbl>, q100 <dbl>,
## #   q101 <dbl>, q102 <dbl>, q103 <dbl>, q104 <dbl>, q105 <dbl>,
## #   q106 <dbl>, ...
```

```
#glimpse(crsc96)
```

Sélectionnons les données les variables qui nous intéressent

Ce choix est basé sur notre sujet d'étude, sur la théorie et sur les travaux empiriques dans le domaine. Il est toujours important de ne pas faire cette selection sur la même base de données.

```
crsc96_small <-
  crsc96 %>%
  select(sexq, region, age, ageq, q1, q2, q3, q4, q44, q95, q96)

crsc96_small
```

```
## # A tibble: 2,859 x 11
##     sexq region   age  ageq    q1    q2    q3    q4   q44   q95   q96
##    <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1     2      9    33     3     1     5     5     5     4     5     1
##  2     2      9    34     3     2     5     4     5     5     5     5
##  3     2      9    56     4     2     2     4     5     5     4     5
##  4     1      9    69     5     1     4     2     4     5     5     2
##  5     1      9    43     3     4     4     4     5     5     4     4
##  6     2      9    28     2     4     5     4     5     5     5     2
##  7     1      9    27     2     2     4     2     4     4     5     4
##  8     1      9    51     4     1     4     4     5     5     4     2
##  9     1      9    41     3     1     5     5     5     4     4     5
## 10     1      9    39     3     4     2     5     5     4     5     1
## # ... with 2,849 more rows
```

Sélectionner les observations

```
crsc96_small_homme <-
  crsc96_small %>%
  filter(sexq == 1 & age >= 35)

# Vérification
freq(crsc96$age)
```

```
## Frequencies
## crsc96$age
## Type: Numeric
##
##               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##          15     49      1.71           1.71      1.71           1.71
##          16     55      1.92           3.64      1.92           3.64
##          17     52      1.82           5.46      1.82           5.46
##          18     64      2.24           7.69      2.24           7.69
##          19     57      1.99           9.69      1.99           9.69
##          20     51      1.78          11.47      1.78          11.47
##          21     48      1.68          13.15      1.68          13.15
##          22     51      1.78          14.94      1.78          14.94
##          23     58      2.03          16.96      2.03          16.96
##          24     54      1.89          18.85      1.89          18.85
##          25     56      1.96          20.81      1.96          20.81
##          26     50      1.75          22.56      1.75          22.56
##          27     46      1.61          24.17      1.61          24.17
```

```
##          28      65      2.27          26.44      2.27          26.44
##          29      73      2.55          29.00      2.55          29.00
##          30      67      2.34          31.34      2.34          31.34
##          31      46      1.61          32.95      1.61          32.95
##          32      62      2.17          35.12      2.17          35.12
##          33      53      1.85          36.97      1.85          36.97
##          34      65      2.27          39.24      2.27          39.24
##          35      58      2.03          41.27      2.03          41.27
##          36      66      2.31          43.58      2.31          43.58
##          37      58      2.03          45.61      2.03          45.61
##          38      73      2.55          48.16      2.55          48.16
##          39      70      2.45          50.61      2.45          50.61
##          40      62      2.17          52.78      2.17          52.78
##          41      55      1.92          54.70      1.92          54.70
##          42      67      2.34          57.05      2.34          57.05
##          43      52      1.82          58.87      1.82          58.87
##          44      67      2.34          61.21      2.34          61.21
##          45      47      1.64          62.85      1.64          62.85
##          46      63      2.20          65.06      2.20          65.06
##          47      39      1.36          66.42      1.36          66.42
##          48      55      1.92          68.35      1.92          68.35
##          49      38      1.33          69.67      1.33          69.67
##          50      41      1.43          71.11      1.43          71.11
##          51      38      1.33          72.44      1.33          72.44
##          52      33      1.15          73.59      1.15          73.59
##          53      24      0.84          74.43      0.84          74.43
##          54      33      1.15          75.59      1.15          75.59
##          55      32      1.12          76.71      1.12          76.71
##          56      32      1.12          77.82      1.12          77.82
##          57      25      0.87          78.70      0.87          78.70
##          58      40      1.40          80.10      1.40          80.10
##          59      43      1.50          81.60      1.50          81.60
##          60      44      1.54          83.14      1.54          83.14
##          61      30      1.05          84.19      1.05          84.19
##          62      31      1.08          85.27      1.08          85.27
##          63      34      1.19          86.46      1.19          86.46
##          64      39      1.36          87.83      1.36          87.83
##          65      32      1.12          88.95      1.12          88.95
##          66      28      0.98          89.93      0.98          89.93
##          67      32      1.12          91.05      1.12          91.05
##          68      27      0.94          91.99      0.94          91.99
##          69      34      1.19          93.18      1.19          93.18
##          70      28      0.98          94.16      0.98          94.16
##          71      25      0.87          95.03      0.87          95.03
##          72      28      0.98          96.01      0.98          96.01
##          73      16      0.56          96.57      0.56          96.57
##          74      11      0.38          96.96      0.38          96.96
##          75      14      0.49          97.45      0.49          97.45
##          76      14      0.49          97.94      0.49          97.94
##          77      13      0.45          98.39      0.45          98.39
##          78       6      0.21          98.60      0.21          98.60
##          79       5      0.17          98.78      0.17          98.78
##          80       4      0.14          98.92      0.14          98.92
##          81       4      0.14          99.06      0.14          99.06
```

```
##          82      3      0.10          99.16      0.10          99.16
##          83      5      0.17          99.34      0.17          99.34
##          84      1      0.03          99.37      0.03          99.37
##          85      3      0.10          99.48      0.10          99.48
##          86      5      0.17          99.65      0.17          99.65
##          87      1      0.03          99.69      0.03          99.69
##          88      2      0.07          99.76      0.07          99.76
##          90      1      0.03          99.79      0.03          99.79
##          95      1      0.03          99.83      0.03          99.83
##          99      5      0.17         100.00      0.17         100.00
##       <NA>       0                                0.00         100.00
##       Total   2859    100.00         100.00    100.00         100.00
```

```
freq(crsc96_small_homme$age)
```

```
## Frequencies
## crsc96_small_homme$age
## Type: Numeric
##
##               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##          35     23      2.86           2.86      2.86           2.86
##          36     27      3.36           6.22      3.36           6.22
##          37     27      3.36           9.58      3.36           9.58
##          38     32      3.98          13.56      3.98          13.56
##          39     35      4.35          17.91      4.35          17.91
##          40     36      4.48          22.39      4.48          22.39
##          41     27      3.36          25.75      3.36          25.75
##          42     34      4.23          29.98      4.23          29.98
##          43     26      3.23          33.21      3.23          33.21
##          44     38      4.73          37.94      4.73          37.94
##          45     20      2.49          40.42      2.49          40.42
##          46     32      3.98          44.40      3.98          44.40
##          47     23      2.86          47.26      2.86          47.26
##          48     19      2.36          49.63      2.36          49.63
##          49     20      2.49          52.11      2.49          52.11
##          50     15      1.87          53.98      1.87          53.98
##          51     15      1.87          55.85      1.87          55.85
##          52     15      1.87          57.71      1.87          57.71
##          53     11      1.37          59.08      1.37          59.08
##          54     16      1.99          61.07      1.99          61.07
##          55     19      2.36          63.43      2.36          63.43
##          56     12      1.49          64.93      1.49          64.93
##          57     16      1.99          66.92      1.99          66.92
##          58     13      1.62          68.53      1.62          68.53
##          59     22      2.74          71.27      2.74          71.27
##          60     15      1.87          73.13      1.87          73.13
##          61     15      1.87          75.00      1.87          75.00
##          62     14      1.74          76.74      1.74          76.74
##          63     17      2.11          78.86      2.11          78.86
##          64     15      1.87          80.72      1.87          80.72
##          65     12      1.49          82.21      1.49          82.21
##          66     15      1.87          84.08      1.87          84.08
##          67     15      1.87          85.95      1.87          85.95
##          68     11      1.37          87.31      1.37          87.31
```

```
##          69     17      2.11          89.43      2.11          89.43
##          70     13      1.62          91.04      1.62          91.04
##          71      9      1.12          92.16      1.12          92.16
##          72     12      1.49          93.66      1.49          93.66
##          73      6      0.75          94.40      0.75          94.40
##          74      4      0.50          94.90      0.50          94.90
##          75      8      1.00          95.90      1.00          95.90
##          76      8      1.00          96.89      1.00          96.89
##          77      4      0.50          97.39      0.50          97.39
##          78      2      0.25          97.64      0.25          97.64
##          79      3      0.37          98.01      0.37          98.01
##          80      2      0.25          98.26      0.25          98.26
##          81      3      0.37          98.63      0.37          98.63
##          82      1      0.12          98.76      0.12          98.76
##          83      3      0.37          99.13      0.37          99.13
##          85      2      0.25          99.38      0.25          99.38
##          86      2      0.25          99.63      0.25          99.63
##          87      1      0.12          99.75      0.12          99.75
##          88      1      0.12          99.88      0.12          99.88
##          90      1      0.12         100.00      0.12         100.00
##        <NA>      0                                0.00         100.00
##       Total    804    100.00         100.00    100.00         100.00
```

Toutes ces étapes peuvent se réduire à:

```
crsc96_small_homme_general <-
  crsc96 %>%
  select(sexq, region, age, ageq, q1, q4, q44, q95) %>%
  filter(sexq == 1 & age >= 35)
```

- Autre fonction qui fait la même chose **subset**

Classe des variables

```
class(crsc96_small$q2)
```

```
## [1] "numeric"
```

```
#freq(crsc96_small$q1)
```

- Allons regarder voir si cette variable est vraiment numérique.

R donne un nom particulier aux types de variables différent des noms que les statistiques leur donne:

Statistiques	R
Quantitative	numeric
Entier (âge)	integer
décimal (taille)	dbl (pour double)
Qualitative	
nominal ou ordinal	factor
Caractères (texte)	character (on ne va beaucoup travailler avec ceci)
logique (Vrai / faux)	logical

Fréquences sur une variable

```
freq(crsc96$q2)
```

```
## Frequencies
## crsc96$q2
## Type: Numeric
##
##               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##           1    516     18.05          18.05     18.05          18.05
##           2    636     22.25          40.29     22.25          40.29
##           3     26      0.91          41.20      0.91          41.20
##           4    894     31.27          72.47     31.27          72.47
##           5    787     27.53         100.00     27.53         100.00
##        <NA>      0                                0.00         100.00
##       Total   2859    100.00         100.00    100.00         100.00
```

- Que faites vous si vous ne voulez pas avoir les pourcentages?

Recodage et création de variables facorielles

- Comme on l'a vu, la variable q1 n'est pas numérique mais qualitative (ordinal ? ou nominal ?)
- La création de nouvelles variables se fait avec la commande `mutate`

```
crsc96_small <-
  crsc96_small %>%
  mutate(q2_new = case_when(
    q1 == 1 ~ "totally agree",
    q1 == 2 ~ "agree somewhat",
    q1 == 3 ~ "DK/NA",
    q1 == 4 ~ "disagree somewhat",
    q1 == 5 ~ "totally disagree"))
```

Quelle est la classe de cette nouvelle variable? Quelle est la fréquence de distribution?

```
class(crsc96_small$q2_new)
```

```
## [1] "character"
```

```
freq(crsc96_small$q2_new)
```

```
## Frequencies
## crsc96_small$q2_new
## Type: Character
##
##                         Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ---------------------- ------ --------- -------------- --------- --------------
##          agree somewhat   1410     49.32          49.32     49.32          49.32
##       disagree somewhat    326     11.40          60.72     11.40          60.72
##                   DK/NA      9      0.31          61.04      0.31          61.04
##           totally agree   1065     37.25          98.29     37.25          98.29
##        totally disagree     49      1.71         100.00      1.71         100.00
##                    <NA>      0                                0.00         100.00
##                   Total   2859    100.00         100.00    100.00         100.00
```

Il faut le changer alors en variable factorielle. On vera comment faire bientôt.

If_else pour créer des variables binaires ou dichotomiques

Supposons que nous voulons scinder la variable age en deux catégorie, alors on peut utiliser la commande `if_else`

if_else(condition, valeur si la condition est vraie, valeur si la condition est fausse)

```
crsc96_small <-
  crsc96_small %>%
  mutate(age2 = if_else(age >= 35, "adulte", "jeune"))

class(crsc96_small$age2)
```

```
## [1] "character"
```

```
freq(crsc96_small$age2)
```

```
## Frequencies
## crsc96_small$age2
## Type: Character
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ------------ ------ --------- -------------- --------- --------------
##       adulte   1737     60.76          60.76     60.76          60.76
##        jeune   1122     39.24         100.00     39.24         100.00
##         <NA>      0                               0.00         100.00
##        Total   2859    100.00         100.00    100.00         100.00
```

Commande case_when pour des cas plus généraux

```
crsc96_small <-
  crsc96_small %>%
  mutate(age4 = case_when(
    age < 20 ~ "adolescent",
    age >= 20 & age < 34 ~ "jeune",
    age >= 35 & age <59 ~ "adulte",
    age >=60 ~ "ainé"
  ))

class(crsc96_small$age4)
```

```
## [1] "character"
```

```
freq(crsc96_small$age4)
```

```
## Frequencies
## crsc96_small$age4
## Type: Character
##
##                   Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ---------------- ------ --------- -------------- --------- --------------
##       adolescent    277     10.07          10.07      9.69           9.69
##           adulte   1168     42.46          52.53     40.85          50.54
```

```
##             ainé    526     19.12          71.65     18.40          68.94
##            jeune    780     28.35         100.00     27.28          96.22
##             <NA>    108                                3.78         100.00
##            Total   2859    100.00         100.00    100.00         100.00
```

Pour le rendre comme une variable catégorielle

```
crsc96_small <-
  crsc96_small %>%
  mutate(age4 = as.factor(age4))

class(crsc96_small$age4)
```

```
## [1] "factor"
```

```
freq(crsc96_small$age4)
```

```
## Frequencies
## crsc96_small$age4
## Type: Factor
##
##                    Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ---------------- ------ --------- -------------- --------- --------------
##       adolescent    277     10.07          10.07      9.69           9.69
##           adulte   1168     42.46          52.53     40.85          50.54
##             ainé    526     19.12          71.65     18.40          68.94
##            jeune    780     28.35         100.00     27.28          96.22
##             <NA>    108                                3.78         100.00
##            Total   2859    100.00         100.00    100.00         100.00
```

Distribution de fréquences et de pourcentage (Chap 2)

```
nombre_sexe <-
  crsc96_small %>%
  count(sexe = sexq)
nombre_sexe
```

```
## # A tibble: 2 x 2
##    sexe     n
##   <dbl> <int>
## 1     1  1361
## 2     2  1498
```

```
nombre_age4 <-
  crsc96_small %>%
  count(age = age4)
```

```
## Warning: Factor `age` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

```
nombre_age4
```

```
## # A tibble: 5 x 2
##   age            n
```

```
##   <fct>      <int>
## 1 adolescent   277
## 2 adulte      1168
## 3 ainé         526
## 4 jeune        780
## 5 <NA>         108
```

Calculer des proportions

```
proportion_sexe <-
  crsc96_small %>%
  count(sexe = sexq) %>%
  mutate(proportion = n / (sum(n)))
proportion_sexe
```

```
## # A tibble: 2 x 3
##    sexe     n proportion
##   <dbl> <int>      <dbl>
## 1     1  1361      0.476
## 2     2  1498      0.524
```

```
# proportion

proportion_age4 <-
  crsc96_small %>%
  count(age4) %>%
  mutate(proportion = n / (sum(n)))
```

```
## Warning: Factor `age4` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

```
proportion_age4
```

```
## # A tibble: 5 x 3
##   age4           n proportion
##   <fct>      <int>      <dbl>
## 1 adolescent   277     0.0969
## 2 adulte      1168     0.409
## 3 ainé         526     0.184
## 4 jeune        780     0.273
## 5 <NA>         108     0.0378
```

Avec le package Summarytools

Les mêmes résultats sont obtenus directement avec freq de summarytools tel qu'on l'a vu précédemment.

```
freq(crsc96$sexq)
```

```
## Frequencies
## crsc96$sexq
## Type: Numeric
##
##               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ------------ ------ --------- -------------- --------- --------------
```

```
##           1   1361     47.60          47.60     47.60          47.60
##           2   1498     52.40         100.00     52.40         100.00
##        <NA>      0                               0.00         100.00
##       Total   2859    100.00         100.00    100.00         100.00
```

```
freq(crsc96$q1)
```

```
## Frequencies
## crsc96$q1
## Type: Numeric
##
##               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##           1   1065     37.25          37.25     37.25          37.25
##           2   1410     49.32          86.57     49.32          86.57
##           3      9      0.31          86.88      0.31          86.88
##           4    326     11.40          98.29     11.40          98.29
##           5     49      1.71         100.00      1.71         100.00
##        <NA>      0                               0.00         100.00
##       Total   2859    100.00         100.00    100.00         100.00
```

```
freq(crsc96$region)
```

```
## Frequencies
## crsc96$region
## Type: Numeric
##
##               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##           0    218      7.63           7.63      7.63           7.63
##           1    270      9.44          17.07      9.44          17.07
##           2    564     19.73          36.80     19.73          36.80
##           3    531     18.57          55.37     18.57          55.37
##           4    211      7.38          62.75      7.38          62.75
##           5    351     12.28          75.03     12.28          75.03
##           6    124      4.34          79.36      4.34          79.36
##           7    117      4.09          83.46      4.09          83.46
##           8    240      8.39          91.85      8.39          91.85
##           9    233      8.15         100.00      8.15         100.00
##        <NA>      0                               0.00         100.00
##       Total   2859    100.00         100.00    100.00         100.00
```

```
freq(crsc96$q44)
```

```
## Frequencies
## crsc96$q44
## Type: Numeric
##
##               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##           1    118      4.13           4.13      4.13           4.13
##           2    414     14.48          18.61     14.48          18.61
##           3     18      0.63          19.24      0.63          19.24
##           4   1293     45.23          64.46     45.23          64.46
##           5   1016     35.54         100.00     35.54         100.00
##        <NA>      0                               0.00         100.00
```

```
##       Total   2859    100.00         100.00    100.00         100.00
```

```
freq(crsc96$q95)
```

```
## Frequencies
## crsc96$q95
## Type: Numeric
##
##               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##           1     66      2.31           2.31      2.31           2.31
##           2    240      8.39          10.70      8.39          10.70
##           3     22      0.77          11.47      0.77          11.47
##           4    605     21.16          32.63     21.16          32.63
##           5   1926     67.37         100.00     67.37         100.00
##        <NA>      0                                0.00         100.00
##       Total   2859    100.00         100.00    100.00         100.00
```

Application

- Créer la variable age au carré nommé `age_square`
- Recoder la variable q2 en trois catégories (`agree`, `dk`, et `disagree`) (variable factorielle)
- Créer une nouvelle variable qui permet de savoir combien de personne sont dans le groupe d'âge [25, 35]
- Créer une variable qui divise l'age en 5 catégories
- Créer la variable age_ecart qui est l'écart de la valeur de l'âge par rapport à la moyenne

```
crsc96_small <-
  crsc96_small %>%
  mutate(age_square = age^2,
         q2_3 = factor(case_when(
           q2 == 1 | q2 == 2 ~ "agree",
           q2 == 3 ~ "dk",
           q2 == 4 | q2 == 5 ~ "disagree")),
         age_groupe = between(age, 25, 34))

class(crsc96_small$q2_3)
```

```
## [1] "factor"
```

```
freq(crsc96_small$age_groupe)
```

```
## Frequencies
## crsc96_small$age_groupe
## Type: Logical
##
##               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##       FALSE   2276     79.61          79.61     79.61          79.61
##        TRUE    583     20.39         100.00     20.39         100.00
##        <NA>      0                                0.00         100.00
##       Total   2859    100.00         100.00    100.00         100.00
```

```
freq(crsc96_small$q2_3)
```

```
## Frequencies
## crsc96_small$q2_3
```

```
## Type: Factor
##
##                  Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## -------------- ------ --------- -------------- --------- --------------
##          agree   1152     40.29          40.29     40.29          40.29
##       disagree   1681     58.80          99.09     58.80          99.09
##             dk     26      0.91         100.00      0.91         100.00
##           <NA>      0                                0.00         100.00
##          Total   2859    100.00         100.00    100.00         100.00
```

PAUSE

Statistiques univariées

- Ces statistiques s'appliquent au variables quantitatives.

La commande summary nous donne une première indication sur l'ensemble des variables de notre base de données. Il faut prêter attention aux variables manquantes. D'où proviennent les données manquantes dans age4

```
summary(crsc96_small)
```

```
##       sexq           region           age             ageq
##  Min.   :1.000   Min.   :0.000   Min.   :15.00   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:2.000   1st Qu.:28.00   1st Qu.:2.000
##  Median :2.000   Median :3.000   Median :39.00   Median :3.000
##  Mean   :1.524   Mean   :3.907   Mean   :41.45   Mean   :3.226
##  3rd Qu.:2.000   3rd Qu.:5.000   3rd Qu.:54.00   3rd Qu.:4.000
##  Max.   :2.000   Max.   :9.000   Max.   :99.00   Max.   :5.000
##        q1              q2              q3              q4
##  Min.   :1.00   Min.   :1.00   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.00   1st Qu.:2.00   1st Qu.:2.000   1st Qu.:4.000
##  Median :2.00   Median :4.00   Median :4.000   Median :5.000
##  Mean   :1.91   Mean   :3.28   Mean   :3.685   Mean   :4.524
##  3rd Qu.:2.00   3rd Qu.:5.00   3rd Qu.:5.000   3rd Qu.:5.000
##  Max.   :5.00   Max.   :5.00   Max.   :5.000   Max.   :5.000
##       q44             q95             q96          q2_new
##  Min.   :1.000   Min.   :1.000   Min.   :1.0   Length:2859
##  1st Qu.:4.000   1st Qu.:4.000   1st Qu.:1.0   Class :character
##  Median :4.000   Median :5.000   Median :2.0   Mode  :character
##  Mean   :3.936   Mean   :4.429   Mean   :2.3
##  3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:4.0
##  Max.   :5.000   Max.   :5.000   Max.   :5.0
##      age2                   age4        age_square         q2_3
##  Length:2859        adolescent: 277   Min.   : 225   agree   :1152
##  Class :character   adulte    :1168   1st Qu.: 784   disagree:1681
##  Mode  :character   ainé      : 526   Median :1521   dk      :  26
##                     jeune     : 780   Mean   :2011
##                     NA's      : 108   3rd Qu.:2916
##                                       Max.   :9801
##  age_groupe
##  Mode :logical
##  FALSE:2276
##  TRUE :583
```

```
##
##
##
```

L'inconvénient, c'est que c'est mal présenté, et ce ne sont pas l'ensemble des variables de notre base de données qui nous concernent. Les informations sur les variables caractères ne sont pas fournies. C'est pourquoi, il faut toujours les transformer en variables factorielles.

Paramètres de position

```
age_moyen <- mean(crsc96_small$age)
age_moyen
```

```
## [1] 41.45261
```

```
age_median <- median(crsc96_small$age)
age_median
```

```
## [1] 39
```

Cette approche n'est pas la bonne car elle nous demande beaucoup de coding (avec la création de plusieurs objets)

Statistiques univariées

La fonction `summarise` permet de calculer l'ensemble des indicateurs dont nous avons besoin. Dans toute étude, il est important de résumé l'information contenue dans les variables pour se faire une première idée.

```
age_position <-
  crsc96_small %>%
  summarise(age_moyen = mean(age),
            age_median = median(age),
            age_Q1 = quantile(age, prob = 0.25),
            age_Q3 = quantile(age, prob = 0.75),
            age_min = min(age))
age_position
```

```
## # A tibble: 1 x 5
##   age_moyen age_median age_Q1 age_Q3 age_min
##       <dbl>      <dbl>  <dbl>  <dbl>   <dbl>
## 1      41.5         39     28     54      15
```

Statistiques univariées: Mode

Il n'y a aucune fonction qui permet de calculer directement le mode. Alors, il faut créer cette fonction soit même ou utiliser celle créer par un autre utilisateur.

```
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

age_position <-
  crsc96_small %>%
```

```
  summarise(age_moyen = mean(age),
            age_median = median(age),
            age_Q1 = quantile(age, prob = 0.25),
            age_Q3 = quantile(age, prob = 0.75),
            age_mode = getmode(age))

age_position
```

```
## # A tibble: 1 x 5
##   age_moyen age_median age_Q1 age_Q3 age_mode
##       <dbl>      <dbl>  <dbl>  <dbl>    <dbl>
## 1      41.5         39     28     54       29
```

```
# Vérification du mode
freq(crsc96_small$age)
```

```
## Frequencies
## crsc96_small$age
## Type: Numeric
##
##               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##          15     49      1.71           1.71      1.71           1.71
##          16     55      1.92           3.64      1.92           3.64
##          17     52      1.82           5.46      1.82           5.46
##          18     64      2.24           7.69      2.24           7.69
##          19     57      1.99           9.69      1.99           9.69
##          20     51      1.78          11.47      1.78          11.47
##          21     48      1.68          13.15      1.68          13.15
##          22     51      1.78          14.94      1.78          14.94
##          23     58      2.03          16.96      2.03          16.96
##          24     54      1.89          18.85      1.89          18.85
##          25     56      1.96          20.81      1.96          20.81
##          26     50      1.75          22.56      1.75          22.56
##          27     46      1.61          24.17      1.61          24.17
##          28     65      2.27          26.44      2.27          26.44
##          29     73      2.55          29.00      2.55          29.00
##          30     67      2.34          31.34      2.34          31.34
##          31     46      1.61          32.95      1.61          32.95
##          32     62      2.17          35.12      2.17          35.12
##          33     53      1.85          36.97      1.85          36.97
##          34     65      2.27          39.24      2.27          39.24
##          35     58      2.03          41.27      2.03          41.27
##          36     66      2.31          43.58      2.31          43.58
##          37     58      2.03          45.61      2.03          45.61
##          38     73      2.55          48.16      2.55          48.16
##          39     70      2.45          50.61      2.45          50.61
##          40     62      2.17          52.78      2.17          52.78
##          41     55      1.92          54.70      1.92          54.70
##          42     67      2.34          57.05      2.34          57.05
##          43     52      1.82          58.87      1.82          58.87
##          44     67      2.34          61.21      2.34          61.21
##          45     47      1.64          62.85      1.64          62.85
##          46     63      2.20          65.06      2.20          65.06
```

```
##          47     39      1.36          66.42      1.36          66.42
##          48     55      1.92          68.35      1.92          68.35
##          49     38      1.33          69.67      1.33          69.67
##          50     41      1.43          71.11      1.43          71.11
##          51     38      1.33          72.44      1.33          72.44
##          52     33      1.15          73.59      1.15          73.59
##          53     24      0.84          74.43      0.84          74.43
##          54     33      1.15          75.59      1.15          75.59
##          55     32      1.12          76.71      1.12          76.71
##          56     32      1.12          77.82      1.12          77.82
##          57     25      0.87          78.70      0.87          78.70
##          58     40      1.40          80.10      1.40          80.10
##          59     43      1.50          81.60      1.50          81.60
##          60     44      1.54          83.14      1.54          83.14
##          61     30      1.05          84.19      1.05          84.19
##          62     31      1.08          85.27      1.08          85.27
##          63     34      1.19          86.46      1.19          86.46
##          64     39      1.36          87.83      1.36          87.83
##          65     32      1.12          88.95      1.12          88.95
##          66     28      0.98          89.93      0.98          89.93
##          67     32      1.12          91.05      1.12          91.05
##          68     27      0.94          91.99      0.94          91.99
##          69     34      1.19          93.18      1.19          93.18
##          70     28      0.98          94.16      0.98          94.16
##          71     25      0.87          95.03      0.87          95.03
##          72     28      0.98          96.01      0.98          96.01
##          73     16      0.56          96.57      0.56          96.57
##          74     11      0.38          96.96      0.38          96.96
##          75     14      0.49          97.45      0.49          97.45
##          76     14      0.49          97.94      0.49          97.94
##          77     13      0.45          98.39      0.45          98.39
##          78      6      0.21          98.60      0.21          98.60
##          79      5      0.17          98.78      0.17          98.78
##          80      4      0.14          98.92      0.14          98.92
##          81      4      0.14          99.06      0.14          99.06
##          82      3      0.10          99.16      0.10          99.16
##          83      5      0.17          99.34      0.17          99.34
##          84      1      0.03          99.37      0.03          99.37
##          85      3      0.10          99.48      0.10          99.48
##          86      5      0.17          99.65      0.17          99.65
##          87      1      0.03          99.69      0.03          99.69
##          88      2      0.07          99.76      0.07          99.76
##          90      1      0.03          99.79      0.03          99.79
##          95      1      0.03          99.83      0.03          99.83
##          99      5      0.17         100.00      0.17         100.00
##        <NA>      0                                0.00         100.00
##       Total   2859    100.00         100.00    100.00         100.00
```

Statistique par groupe

Nous pouvons aussi regarder ces données selon le sexe des individus. C'est dire voir les statistiques pour les femmes et pour les hommes.

```
age_position_sexe <-
  crsc96_small %>%
  group_by(sexq) %>%
  summarise(age_moyen = mean(age),
            age_median = median(age),
            age_Q1 = quantile(age, prob = 0.25),
            age_Q3 = quantile(age, prob = 0.75),
            age_mode = getmode(age))

age_position_sexe
```

```
## # A tibble: 2 x 6
##    sexq age_moyen age_median age_Q1 age_Q3 age_mode
##   <dbl>     <dbl>      <dbl>  <dbl>  <dbl>    <dbl>
## 1     1      40.8         39     27     52       44
## 2     2      42.0         40     28     55       38
```

EXERCICE

Calculer les paramètres de dispersion de la variable age et commenter.

Données manquantes

- Qu'arrive-t-il si l'âge avait des données manquantes?
- Si vos données contiennent des données manquantes, les statistiques univariées ne vont pas fonctionner. Vous devez lui dire explicitement de les enlever avant de calculer les statistiques

Créons la variable **age_avec_manquant** qui a des valeurs manquantes pour tout ceux qui ont 70 ans ou plus

```
crsc96_small <-
  crsc96_small %>%
  mutate(age_avec_manquant = if_else(age < 70, age, NA_real_))

freq(crsc96_small$age_avec_manquant)
```

```
## Frequencies
## crsc96_small$age_avec_manquant
## Type: Numeric
##
##               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##          15     49      1.84           1.84      1.71           1.71
##          16     55      2.06           3.90      1.92           3.64
##          17     52      1.95           5.86      1.82           5.46
##          18     64      2.40           8.26      2.24           7.69
##          19     57      2.14          10.40      1.99           9.69
##          20     51      1.91          12.31      1.78          11.47
##          21     48      1.80          14.11      1.68          13.15
##          22     51      1.91          16.03      1.78          14.94
##          23     58      2.18          18.21      2.03          16.96
##          24     54      2.03          20.23      1.89          18.85
##          25     56      2.10          22.33      1.96          20.81
```

```
##          26     50      1.88          24.21      1.75          22.56
##          27     46      1.73          25.94      1.61          24.17
##          28     65      2.44          28.38      2.27          26.44
##          29     73      2.74          31.12      2.55          29.00
##          30     67      2.52          33.63      2.34          31.34
##          31     46      1.73          35.36      1.61          32.95
##          32     62      2.33          37.69      2.17          35.12
##          33     53      1.99          39.68      1.85          36.97
##          34     65      2.44          42.12      2.27          39.24
##          35     58      2.18          44.29      2.03          41.27
##          36     66      2.48          46.77      2.31          43.58
##          37     58      2.18          48.95      2.03          45.61
##          38     73      2.74          51.69      2.55          48.16
##          39     70      2.63          54.32      2.45          50.61
##          40     62      2.33          56.64      2.17          52.78
##          41     55      2.06          58.71      1.92          54.70
##          42     67      2.52          61.22      2.34          57.05
##          43     52      1.95          63.18      1.82          58.87
##          44     67      2.52          65.69      2.34          61.21
##          45     47      1.76          67.45      1.64          62.85
##          46     63      2.36          69.82      2.20          65.06
##          47     39      1.46          71.28      1.36          66.42
##          48     55      2.06          73.35      1.92          68.35
##          49     38      1.43          74.77      1.33          69.67
##          50     41      1.54          76.31      1.43          71.11
##          51     38      1.43          77.74      1.33          72.44
##          52     33      1.24          78.98      1.15          73.59
##          53     24      0.90          79.88      0.84          74.43
##          54     33      1.24          81.12      1.15          75.59
##          55     32      1.20          82.32      1.12          76.71
##          56     32      1.20          83.52      1.12          77.82
##          57     25      0.94          84.46      0.87          78.70
##          58     40      1.50          85.96      1.40          80.10
##          59     43      1.61          87.58      1.50          81.60
##          60     44      1.65          89.23      1.54          83.14
##          61     30      1.13          90.35      1.05          84.19
##          62     31      1.16          91.52      1.08          85.27
##          63     34      1.28          92.79      1.19          86.46
##          64     39      1.46          94.26      1.36          87.83
##          65     32      1.20          95.46      1.12          88.95
##          66     28      1.05          96.51      0.98          89.93
##          67     32      1.20          97.71      1.12          91.05
##          68     27      1.01          98.72      0.94          91.99
##          69     34      1.28         100.00      1.19          93.18
##        <NA>    195                                6.82         100.00
##       Total   2859    100.00         100.00    100.00         100.00
```

Qu'est-ce qui s'est passé. En fait, **age_avec_manquant** comporte des valeurs manquantes. Il faut indiquer dans le calcul des statistiques univariées qu'il y a des valeurs manquantes, et qu'il faut les enlever avant de calculer la moyenne, ou toute autre statistique.

```
age_avec_manquant_position <-
  crsc96_small %>%
  summarise(age_moyen = mean(age_avec_manquant, na.rm = TRUE),
            age_median = median(age_avec_manquant, na.rm = TRUE),
```

```
            age_Q1 = quantile(age_avec_manquant, prob = 0.25, na.rm = TRUE),
            age_Q3 = quantile(age_avec_manquant, prob = 0.75, na.rm = TRUE),
            age_mode = getmode(age))

age_avec_manquant_position
```

```
## # A tibble: 1 x 5
##   age_moyen age_median age_Q1 age_Q3 age_mode
##       <dbl>      <dbl>  <dbl>  <dbl>    <dbl>
## 1      39.0         38     27     50       29
```

Quel est le problème qui se pose quand des informations sont manquantes. Peut-on faire confiance aux résultats?

Pouvons-nous calculer les statistiques univariées sur des variables qualitatives?

```
q2_position <-
  crsc96_small %>%
  summarise(q2_moyen = mean(q2, na.rm = TRUE),
            q2_median = median(q2, na.rm = TRUE),
            q2_Q1 = quantile(q2, prob = 0.25, na.rm = TRUE),
            q2_Q3 = quantile(q2, prob = 0.75, na.rm = TRUE),
            q2_mode = getmode(q2))

q2_position
```

```
## # A tibble: 1 x 5
##   q2_moyen q2_median q2_Q1 q2_Q3 q2_mode
##      <dbl>     <dbl> <dbl> <dbl>   <dbl>
## 1     3.28         4     2     5       4
```

Pourquoi ça fonctionne? Comment interprétez-vous ces résultats?

Qu'en est-il de la variable sexe (sexq)

```
sexq_position <-
  crsc96_small %>%
  summarise(sexq_moyen = mean(sexq, na.rm = TRUE),
            sexq_median = median(sexq, na.rm = TRUE),
            sexq_Q1 = quantile(sexq, prob = 0.25, na.rm = TRUE),
            sexq_Q3 = quantile(sexq, prob = 0.75, na.rm = TRUE),
            sexq_mode = getmode(sexq))

sexq_position
```

```
## # A tibble: 1 x 5
##   sexq_moyen sexq_median sexq_Q1 sexq_Q3 sexq_mode
##        <dbl>       <dbl>   <dbl>   <dbl>     <dbl>
## 1       1.52           2       1       2         2
```

Pour une variable dichotomique, seule la moyenne a un sens.

Remarques

1. Tous les objets que vous créez, vous pouvez les manipuler à votre guise

2. Les variables que vous créez, vous pouvez les réutiliser juste après
3. Interprétations des résultats

PAUSE

Représentation graphique pour les distributions univariées

Introduction

- Les graphiques nous permettent de répondre à plusieurs types de questions :
  - Quelle est la distribution d'une variable?
  - Est-ce que les filles ont plus tendances à vivre dans un type particulier de structure familiale?
  - Comment est-ce que la structure de la famille affecte la santé des enfants?
  - Est-ce qu'il existe une association entre les attitudes envers la violence conjugale et le niveau de scolarisation (données dhs_ipv)
  - Cette relation est-elle positive? négative? ou nulle?

Type de graphiques pour les distributions univariées

- Dépend en général du type de variable (qualitative ou quantitative) et du nombre de variable
- Graphiques pour representer une seule variable:

Type de variables	Une seule variable
Qualitative	Diagramme de barre (diagramme en bâton)
	Diagramme circulaire
	Carte (map)
Quantitative	Histogramme (geom_histogram)
	Diagramme de quartile (boîte à moustaches)

- On peut aussi utiliser les graphiques pour representer l'association entre deux variables. On verra cela plus tard

ggplot

Forme générale

- La forme générale d'un code de graphique est le suivant:

```
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
```

1. **ggplot** spécifie que vous utiliser la commande ggplot. C'est à ce niveau que vous spécifier les données que vous voulez utiliser.

- Ce n'est pas toujours obligatoire si vous utilisez plus d'une base de données.

2. **geom_function**, contient plusieurs fonctions pour spécifier le type de graphique que vous voulez faire. Le type de graphique indique le nombre de paramètres à inclure.

- Exemples: geom_histogram() pour les **histogrammes**
- geom_point() pour les **diagrammes de dispertions**,
- geom_barplot() pour les **diagrammes de barre**.
- La liste complète est ici: https://ggplot2.tidyverse.org/reference/

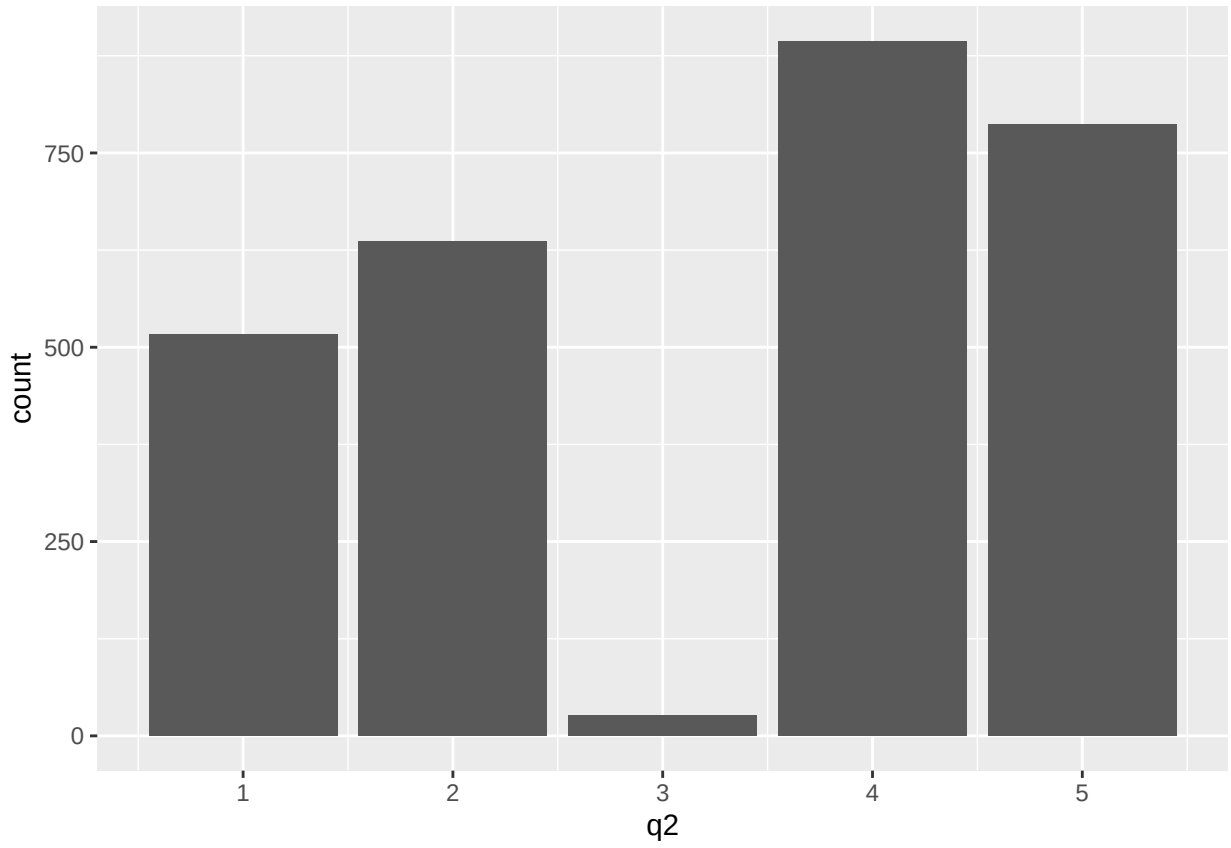
3. **aes** pour aestetics indique le nombre de paramètres à passer à la fonction **geom_function**. Il permet également de spécifier des informations sur le graphique.

Exemples: Visualiser la distribution univariée

Pour les variables qualitatives (nominales et ordinales)

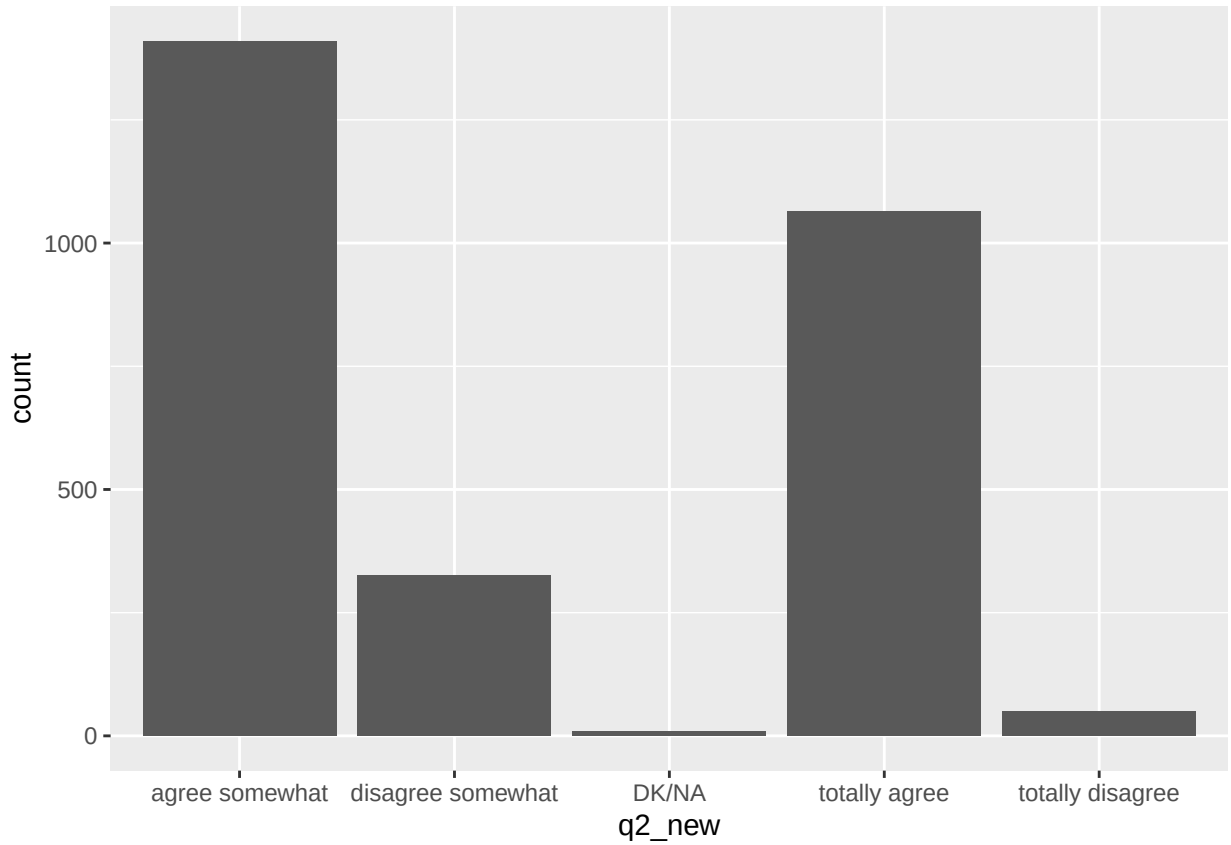
1. Diagramme de barres

```
ggplot(crsc96_small) +
  geom_bar(aes(x = q2))
```



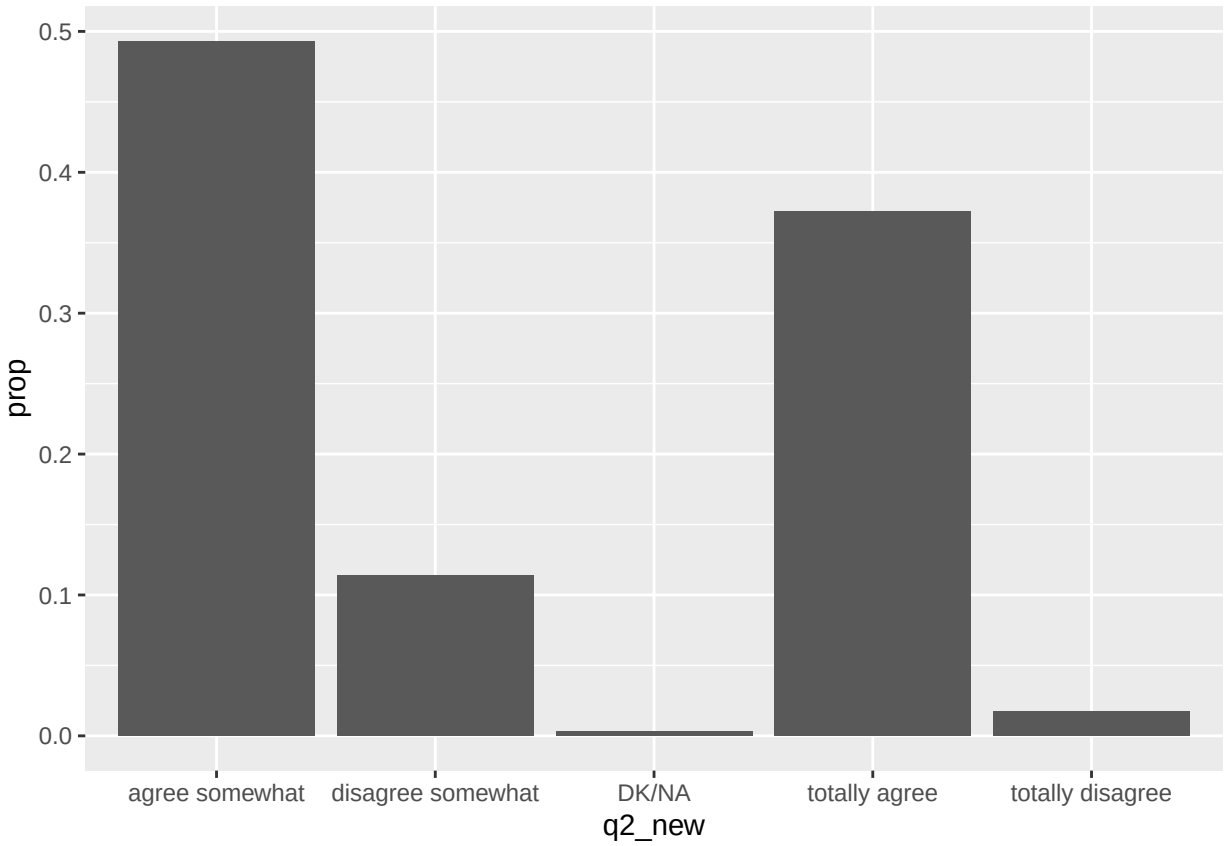
Mieux avec q2_new

```
ggplot(crsc96_small) +
  geom_bar(aes(x = q2_new))
```



- Il faut toujours privilégier les distributions en pourcentages

```
ggplot(crsc96_small) +
  geom_bar(aes(x = q2_new, ..prop.., group = 1))
```



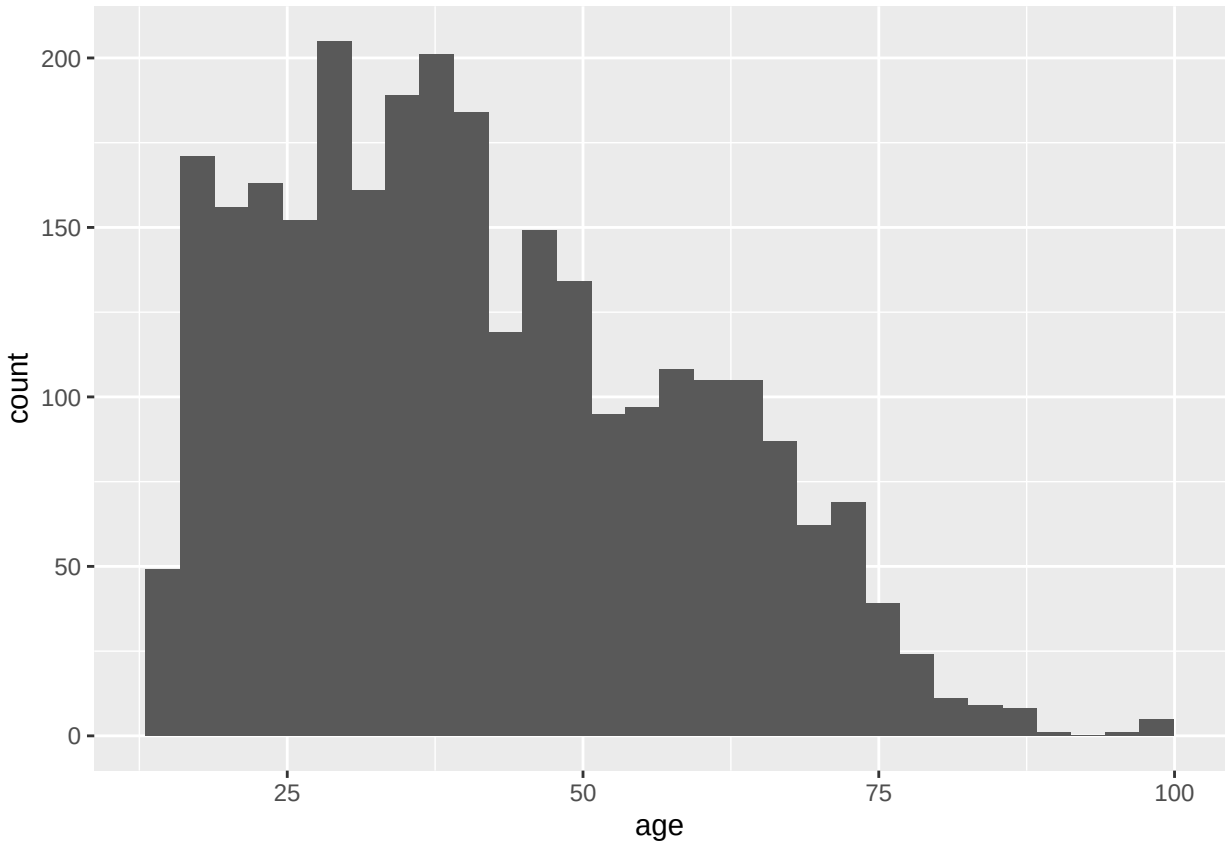
- Diagramme circulaire

Pour les variables quantitatives (intervalle/ratio)

1. Histogramme

```
ggplot(crsc96_small) +
  geom_histogram(aes(x = age))
```

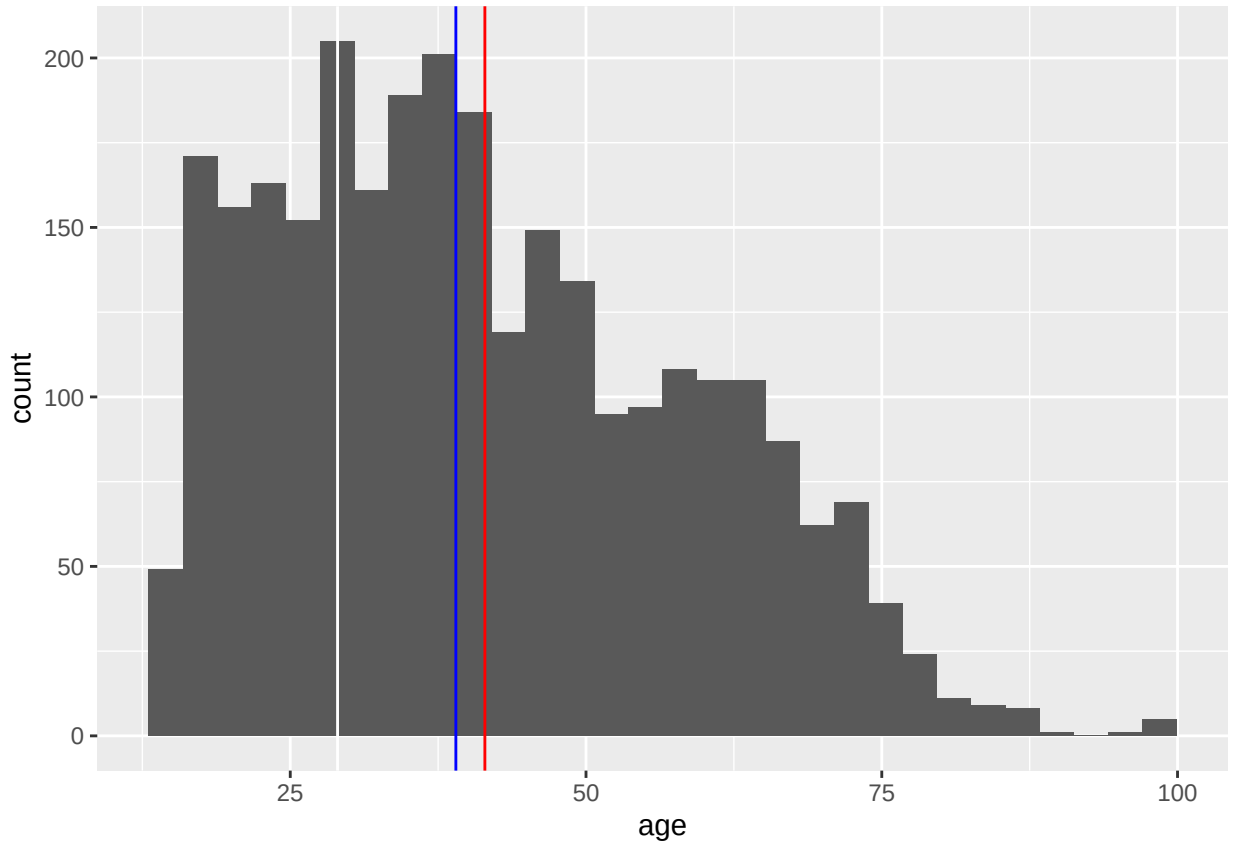
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



On peut ajouter dans ce graphique les informations sur la moyenne, la médiane et le mode

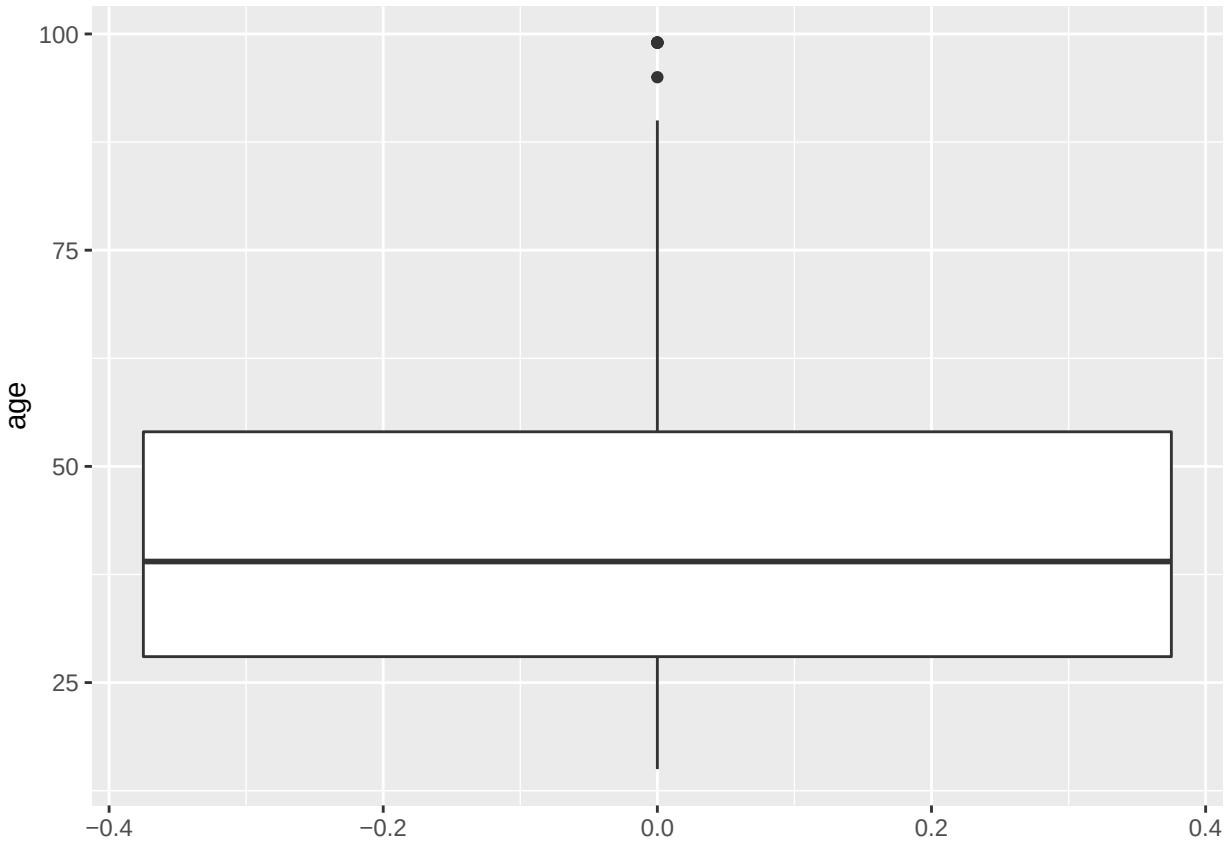
```
ggplot(crsc96_small) +
  geom_histogram(aes(x = age)) +
  geom_vline(aes(xintercept = mean(age)), color = "red") +
  geom_vline(aes(xintercept = median(age)), color = "blue") +
  geom_vline(aes(xintercept = getmode(age)), color = "white")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



2. Diagramme de quartile

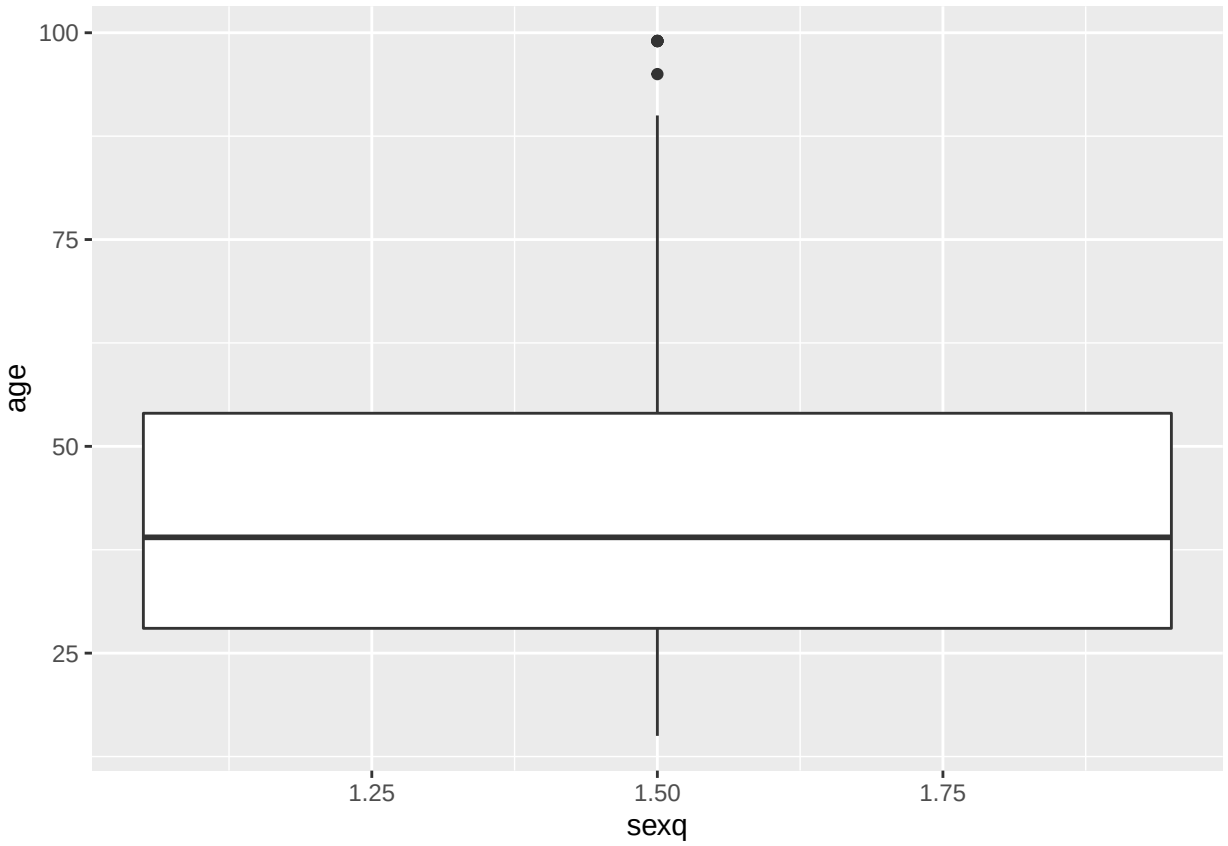
```
ggplot(crsc96_small) +
  geom_boxplot(aes(y = age))
```



Il est plus intéressant si on le calcule pour différent groupes

```
ggplot(crsc96_small) +
  geom_boxplot(aes(x = sexq, y = age))
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



Voici une bonne extension de ce qu'on a vu. Limitez-vous uniquement à ce qu'on a vu: les statistiques univariées. Ce site est magique. Une ressource incontournable pour vous dès maintenant.

http://larmarange.github.io/analyse-R/graphiques-bivaries-ggplot2.html

Et l'extension http://larmarange.github.io/analyse-R/etendre-ggplot2.html

N'oublions pas pour finir la feuille de tricherie

https://thinkr.fr/pdf/ggplot2-french-cheatsheet.pdf