

Le projet R forces

Méziane Cherif

Olivier Cailloux

Contexte

L'assemblée nationale a constitué en 2019 une commission d'enquête concernant « la situation, les missions et les moyens des forces de sécurité ». À cette occasion, elle a interrogé par écrit des agents de la police nationale et des polices municipales, militaires de la gendarmerie nationale et réservistes. Les réponses ont été publiées sur le site de l'assemblée nationale, après anonymisation.

La France a été épinglée ces dernières années à plusieurs reprises par des organisations internationales pour des faits préoccupants de violences policières. Il nous a semblé intéressant dans ce contexte d'examiner si un lien peut être établi entre les conditions socio-économiques des départements d'exercice des forces de sécurité et les réponses données par ces dernières concernant une volonté de répression plus ou moins accrue. (Nous n'avons bien sûr pas de prétention d'apporter de grandes réponses dans le cadre de cet exercice limité de statistique, et encore moins concernant un quelconque établissement de lien de causalité, mais ces données nous ont semblé mériter investigation.)

Nous utilisons également les « Principaux indicateurs sur les revenus et la pauvreté aux niveaux national et local en 2019 » du dispositif Fichier localisé social et fiscal (Filosofi) publié par l'INSEE (présentation ici, téléchargement ici). Des données plus récentes existent mais nous utilisons les données reflétant la réalité au moment des réponses des forces de sécurité.

Mise en place

Chargeons quelques packages utiles.

```
library(conflicted)
conflicts_prefer(dplyr::filter)
library(tidyverse)
```

Téléchargeons les réponses des forces de sécurité, ou vérifions leur conformité si elles sont déjà présentes à l'aide du hash MD5 indiqué sur le site sus-mentionné.

```
answers_url <- paste0(
  "https://data.assemblee-nationale.fr/",
  "static/openData/repository/CONSULTATIONS_CITOYENNES/",
  "MOYENS_DES_FORCES_DE_SECURITE/Moyens-des-forces-de-securite.csv"
)
md5_expected <- "261b4244cc2e9ffcd54ff9a6bec0a0ac"
if (file.exists("Réponses original.csv")) {
  md5_observed <- tools::md5sum("Réponses original.csv")
} else {
  md5_observed <- 0L
}
if (md5_observed != md5_expected) {
  downloaded_return <- download.file(answers_url, "Réponses original.csv", mode = "wb")
  stopifnot(identical(downloaded_return, 0L))
}
```

```

}
md5_observed <- tools::md5sum("Réponses original.csv")
stopifnot(md5_observed == md5_expected)

```

Convertissons en UTF8.

```

input_original <- readLines("Réponses original.csv")
input_converted <- iconv(input_original, from = "WINDOWS-1252", to = "UTF8")
writeLines(input_converted, "Réponses.csv")

```

Téléchargeons de même les données sur les revenus et la pauvreté.

```

zip_file_name <- "base-cc-filosofi-2019_CSV.zip"
filosofi_url <- paste0("https://www.insee.fr/fr/statistiques/fichier/6036902/", zip_file_name)
if (!file.exists(zip_file_name)) {
  downloaded_return <- download.file(filosofi_url, zip_file_name, mode = "wb")
  stopifnot(identical(downloaded_return, 0L))
}

to_extract <- c("cc_filosofi_2019_DEP.csv", "meta_cc_filosofi_2019_DEP.csv")
if (!all(file.exists(to_extract))) {
  unzip(zip_file_name, files = to_extract)
}

```

Lecture des données

Réponses

Lisons les réponses des forces de sécurité.

```

answers <- read_delim("Réponses.csv",
  delim = ";", locale = locale(decimal_mark = ","),
  show_col_types = FALSE, name_repair = "minimal"
)
col_renaming <- read_csv("Colonnes.csv", show_col_types = FALSE)
stopifnot(all.equal(colnames(answers), col_renaming[["Nom original"]]))
colnames(answers) <- col_renaming[["Nouveau nom"]]
answers

```

```

## # A tibble: 13,735 x 73
##   rep d_start d_end you  you_other fct  fct_other belong belong_other dept
##   <dbl> <chr>  <chr> <chr> <chr>    <chr> <chr>    <chr> <chr>
## 1     3 22/05/~ 22/0~ Fonc~ <NA>    Dans~ <NA>    Aux c~ <NA>    08 --
## 2     4 22/05/~ 22/0~ <NA> <NA>    <NA> <NA>    <NA> <NA>
## 3     5 23/05/~ 23/0~ Fonc~ <NA>    Dans~ <NA>    Aux c~ <NA>    95 --
## 4     6 24/05/~ 24/0~ Fonc~ <NA>    Dans~ <NA>    La ge~ <NA>    11 --
## 5     7 24/05/~ 24/0~ Mili~ <NA>    Dans~ <NA>    La ge~ <NA>    74 --
## 6     8 27/05/~ 27/0~ <NA> <NA>    <NA> <NA>    <NA> <NA>
## 7     9 28/05/~ 28/0~ Fonc~ <NA>    Dans~ <NA>    Autre DSPAP 75 --
## 8    10 28/05/~ 28/0~ <NA> <NA>    <NA> <NA>    <NA> <NA>
## 9    11 29/05/~ 29/0~ Mili~ <NA>    Dans~ <NA>    La ge~ <NA>    17 --
## 10   12 29/05/~ 29/0~ Mili~ <NA>    Dans~ <NA>    La ge~ <NA>    58 --
## # i 13,725 more rows
## # i 63 more variables: works <chr>, task_1 <chr>, task_2 <chr>, impr <chr>,
## #   penal <chr>, penal_1 <chr>, penal_2 <chr>, penal_3 <chr>, agemaj <chr>,

```

```
## #   agemaj_other <chr>, hurt <chr>, hurt_then <chr>, prot <chr>,
## #   prot_adeq <chr>, prot_ext <chr>, train <chr>, train_ext <chr>,
## #   train_suff <chr>, train_days_2016 <dbl>, train_days_2017 <dbl>,
## #   train_days_2018 <dbl>, hab_1 <chr>, hab_2 <chr>, hab_3 <chr>, ...
```

Vérifions que les décimales sont lues correctement et que nous disposons du nombre de contributions annoncé sur le site ministériel.

```
stopifnot(answers |> filter(rep == 9) |> pull(train_days_2017) == 2.5)
stopifnot(nrow(answers) == 13735)
```

Revenus et pauvreté

Lisons maintenant les données économiques.

```
revenues_poverty <- read_delim("cc_filosofi_2019_DEP.csv",
  delim = ";", locale = locale(decimal_mark = ","),
  show_col_types = FALSE
)
revenues_poverty
```

```
## # A tibble: 101 x 28
##   CODGEO NBMENFISC19 NBPERSMENFISC19 MED19 PIMP19 TP6019 TP60AGE119 TP60AGE219
##   <chr>      <dbl>          <dbl> <dbl> <dbl> <dbl>      <dbl>      <dbl>
## 1 01          264074          629120 23490  59.6  10.7        15.4        12.4
## 2 02          223635          513278 19880  49.5  18.4         30         23.4
## 3 03          158967          326379 20570  49    15.4        24.9        19.8
## 4 04           74092          154195 20690  51.5  16.6        26.1        21.7
## 5 05           64688          134672 21020  54.2  13.9        21.6         17
## 6 06          527841          1109491 22300  60.4  15.8        22.2        18.1
## 7 07          143925          313991 21010  50.9  14.3        21.5        18.1
## 8 08          117854          260778 19840  47.2  18.6        30.6        22.8
## 9 09           70184          146066 20010  46.7  17.9        27.7        23.2
## 10 10          132913          290472 20580  52.6  16.3        27.8        21.5
## # i 91 more rows
## # i 20 more variables: TP60AGE319 <dbl>, TP60AGE419 <dbl>, TP60AGE519 <dbl>,
## #   TP60AGE619 <dbl>, TP60TOL119 <dbl>, TP60TOL219 <dbl>, PACT19 <dbl>,
## #   PTSA19 <dbl>, PCH019 <dbl>, PBEN19 <dbl>, PPEN19 <dbl>, PPAT19 <dbl>,
## #   PPSOC19 <dbl>, PPFAM19 <dbl>, PPMINI19 <dbl>, PPLOGT19 <dbl>,
## #   PIMPOT19 <dbl>, D119 <dbl>, D919 <dbl>, RD19 <dbl>
```

Vérifions que le revenu médian et le taux de pauvreté de l'Ain sont ceux indiqués sur le site.

```
ain <- revenues_poverty |> filter(CODGEO == "01")
stopifnot(ain |> pull(MED19) == 23490)
stopifnot(ain |> pull(TP6019) == 10.7)
```

Traitement des données

Extrayons le premier mot de la colonne `dept` pour obtenir le code de département (on vérifie avec une réponse donnée que la conversion a fonctionné). Notons que les départements corses ne s'encodent pas comme des nombres, donc ce code doit être de type chaîne de caractères.

Transformons également l'âge donné en nombre entier.

```
stopifnot(answers |> filter(rep == 3) |> pull(dept) == "08 - ARDENNES")
answers <- mutate(answers, dept_nb = str_extract(dept, "[0-9AB]+"), .after = dept)
```

```
stopifnot(answers |> filter(rep == 3) |> pull(dept_nb) == "08")

answers <- answers |>
  filter(!is.na(agemaj)) |>
  filter(agemaj != "Autre")
stopifnot(all(str_detect(answers$agemaj, "[0-9]+ ans$")))
answers <- mutate(answers, agemaj = as.integer(str_extract(agemaj, "[0-9]+")))
```

Croisement des données

Nous pouvons maintenant joindre les données économiques aux réponses des forces de sécurité.

```
data <- left_join(answers, revenues_poverty, by = c("dept_nb" = "CODGEO"))
data

## # A tibble: 7,593 x 101
##   rep d_start d_end you you_other fct fct_other belong belong_other dept
##   <dbl> <chr>  <chr> <chr> <chr>    <chr> <chr>    <chr>  <chr>    <chr>
## 1     9 28/05/~ 28/0~ Fonc~ <NA>    Dans~ <NA>    Autre DSPAP      75 --
## 2    14 29/05/~ 29/0~ Mili~ <NA>    Dans~ <NA>    La ge~ <NA>      25 --
## 3    18 31/05/~ 31/0~ Fonc~ <NA>    Dans~ <NA>    Autre DOPC      75 --
## 4    19 31/05/~ 31/0~ Fonc~ <NA>    Dans~ <NA>    La di~ <NA>      75 --
## 5    20 31/05/~ 31/0~ Fonc~ <NA>    <NA> <NA>    Autre police muni~ 67 --
## 6    22 01/06/~ 01/0~ Fonc~ <NA>    <NA> <NA>    Autre Police muni~ 42 --
## 7    27 03/06/~ 03/0~ Fonc~ <NA>    <NA> <NA>    Autre Police Muni~ 79 --
## 8    28 03/06/~ 03/0~ Fonc~ <NA>    <NA> <NA>    Autre fonction pu~ 01 --
## 9    36 03/06/~ 03/0~ Fonc~ <NA>    <NA> <NA>    Autre <NA>      74 --
## 10   37 03/06/~ 03/0~ Fonc~ <NA>    <NA> <NA>    La di~ <NA>      95 --
## # i 7,583 more rows
## # i 91 more variables: dept_nb <chr>, works <chr>, task_1 <chr>, task_2 <chr>,
## #   impr <chr>, penal <chr>, penal_1 <chr>, penal_2 <chr>, penal_3 <chr>,
## #   agemaj <int>, agemaj_other <chr>, hurt <chr>, hurt_then <chr>, prot <chr>,
## #   prot_adeq <chr>, prot_ext <chr>, train <chr>, train_ext <chr>,
## #   train_suff <chr>, train_days_2016 <dbl>, train_days_2017 <dbl>,
## #   train_days_2018 <dbl>, hab_1 <chr>, hab_2 <chr>, hab_3 <chr>, ...
write_csv(data, "Données fusionnées.csv", na = "")
```

Idées (brouillon)

Nous pourrions analyser le lien entre la pauvreté du lieu d'exercice et l'âge souhaité des mineurs (surtout pauvreté des plus jeunes) ; le manque d'effectif ; le temps de formation ; le nombre de jours supplémentaires impayés et le manque d'effectif ; l'appréciation de la mobilité ; la volonté d'un cadre plus sûr ; la forfaitisation de sanctions...

Sélection

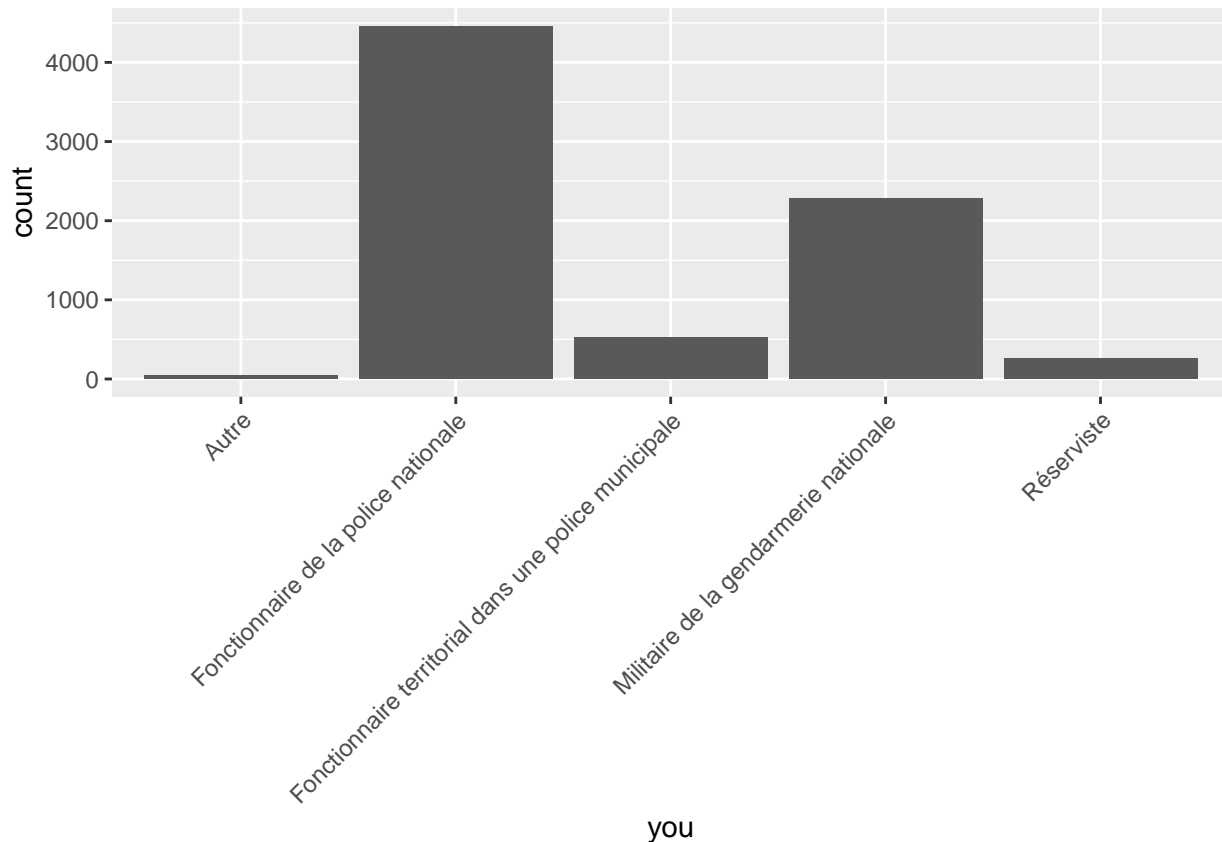
Considérons le taux de pauvreté dans leur département et l'âge souhaité pour traitement comme un majeur. On ne retient en outre que les enregistrements où ces variables sont toutes renseignées.

```
subset <- data |>
  select("you", "TP6019", "agemaj") |>
  drop_na()
```

Comparaison de sous-groupes

Les nombres de réponses par types de répondants diffèrent beaucoup. Considérons les deux types de répondants avec le plus de réponses.

```
subset |> ggplot(aes(x = you)) +  
  geom_bar() +  
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
highest_answers_type <- subset |>  
  select("you") |>  
  group_by(you) |>  
  count() |>  
  ungroup() |>  
  slice_max(n, n = 2) |>  
  pull(you)  
highest_answers_type
```

```
## [1] "Fonctionnaire de la police nationale"  
## [2] "Militaire de la gendarmerie nationale"
```

```
indicators_by_type <- subset |>  
  filter(you %in% highest_answers_type) |>  
  select(you, agemaj) |>  
  group_by(you) |>  
  summarise(mu = mean(agemaj), "s'" = sd(agemaj), n = n())  
indicators_by_type
```

```
## # A tibble: 2 x 4
```

```
##      you                                mu  `s'`      n
##      <chr>                             <dbl> <dbl> <int>
## 1 Fonctionnaire de la police nationale  14.4  1.83  4458
## 2 Militaire de la gendarmerie nationale  14.7  1.63  2286
```

Soit $X^{(1)}$ la variable aléatoire représentant l'âge souhaité pour traitement comme un majeur pour le premier type de répondant et $X^{(2)}$ pour le second type de répondant. Définissons μ_j et σ_j les moyennes et écart-types de $X^{(j)}$ respectivement ($j \in \{1, 2\}$). Notons $X_i^{(j)}$ les observations correspondantes ($j \in \{1, 2\}, i \in \{1, \dots, n_j\}$). On suppose les $X_i^{(j)}$ indépendantes et identiquement distribuées selon $X^{(j)}$.

Avec l'approximation normale (largement valable vu le nombre de réponses), on a $X^{(j)} \approx \mathcal{N}(\mu_j, \sigma_j^2)$ donc $\overline{X^{(j)}} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_i^{(j)} \approx \mathcal{N}(\mu_j, \frac{\sigma_j^2}{n_j})$ et $\overline{X^{(1)}} - \overline{X^{(2)}} \approx \mathcal{N}(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$.

Définissons notre hypothèse nulle H_0 (que nous tentons de réfuter) comme l'égalité des moyennes : $\mu_1 = \mu_2$. Sous H_0 , $\frac{\overline{X^{(1)}} - \overline{X^{(2)}}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx \mathcal{N}(0, 1)$. Avec le Théorème de Slutsky et de continuité, en notant $S'^{(j)2}$ la variance

empirique corrigée des observations $X_i^{(j)}$ et en définissant $Z = \frac{\overline{X^{(1)}} - \overline{X^{(2)}}}{\sqrt{\frac{S'^{(1)2}}{n_1} + \frac{S'^{(2)2}}{n_2}}}$, on obtient $Z \approx \mathcal{N}(0, 1)$.

Notons W la région critique et \overline{W} son complémentaire.

- Avec un risque de première espèce à 5 %, on a $P(Z \in W) = 0.05$ pour $\overline{W} \approx [-1.96, 1.96]$.
- Avec un risque de première espèce à 10 %, on a $P(Z \in W) = 0.1$ pour $\overline{W} \approx [-1.64, 1.64]$.

```
z <- (indicators_by_type$mu[1] - indicators_by_type$mu[2]) / +
  sqrt(
    indicators_by_type[["s'"]][1]^2 / indicators_by_type$n[1] +
    indicators_by_type[["s'"]][2]^2 / indicators_by_type$n[2]
  )
```

Nous observons $z = -6.47 \in W$ et rejettons donc allègrement H_0 : les deux moyennes semblent différentes (à un degré de confiance très élevé, p-value de 9.5e-11).

Voyons ce qu'en pense le test de Welch de R.

```
series1 <- subset |>
  filter(you == highest_answers_type[1]) |>
  pull(agemaj)
series2 <- subset |>
  filter(you == highest_answers_type[2]) |>
  pull(agemaj)
t.test(x = series1, y = series2, var.equal = FALSE)
```

```
##
## Welch Two Sample t-test
##
## data: series1 and series2
## t = -6.4746, df = 5112.4, p-value = 1.04e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.3691140 -0.1975384
## sample estimates:
## mean of x mean of y
## 14.44415 14.72747
```

Ceci confirme nos résultats (bien que le test de Welch n'utilise pas la même distribution approchée que notre approximation par gaussienne comme vu en cours, la différence est extrêmement faible, vu le nombre de nos observations, d'où le fait que la p-value obtenue par R soit du même ordre que la nôtre).

Notons que nous n'avons pas testé l'égalité des variances, nous avons simplement évité de supposer leur égalité, ce qui est plus robuste, ne requiert pas un tel test (controversé dans la littérature), et ne change pas le résultat étant donné que la puissance de notre test est déjà très largement suffisante pour rejeter l'hypothèse nulle.

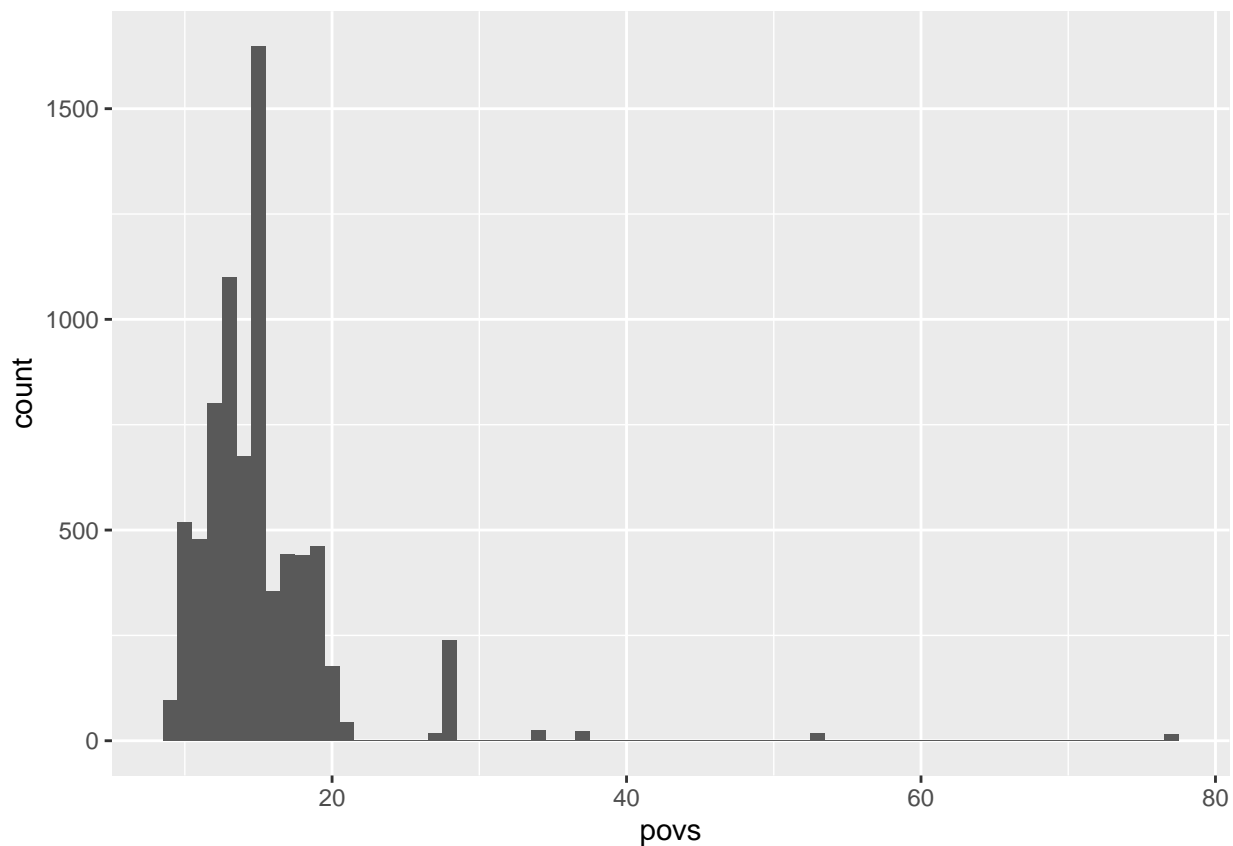
Nous concluons que les répondants de la police nationale et ceux de la gendarmerie nationale ne semblent pas avoir le même âge souhaité pour traitement comme un majeur. La différence est statistiquement très significative, mais il faut noter que la signification pratique de cette différence est très faible, vu la très faible différence observée (en fait elle est statistiquement significative uniquement grâce à notre très grand nombre d'observations).

Lien entre le taux de pauvreté et l'âge souhaité

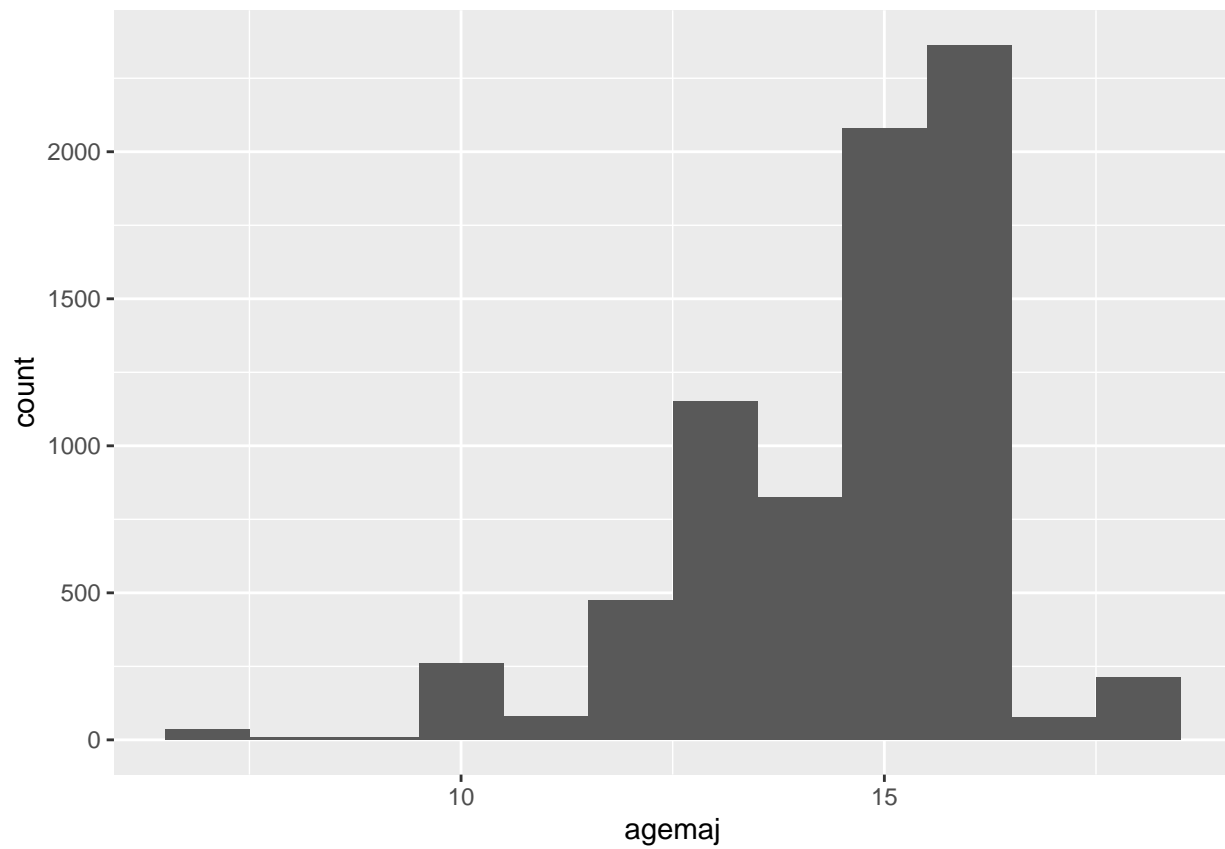
Voyons si le taux de pauvreté dans le département est indépendant de l'âge souhaité pour traitement comme un majeur.

Coupons d'abord les deux séries d'observations en classes d'effectifs proches.

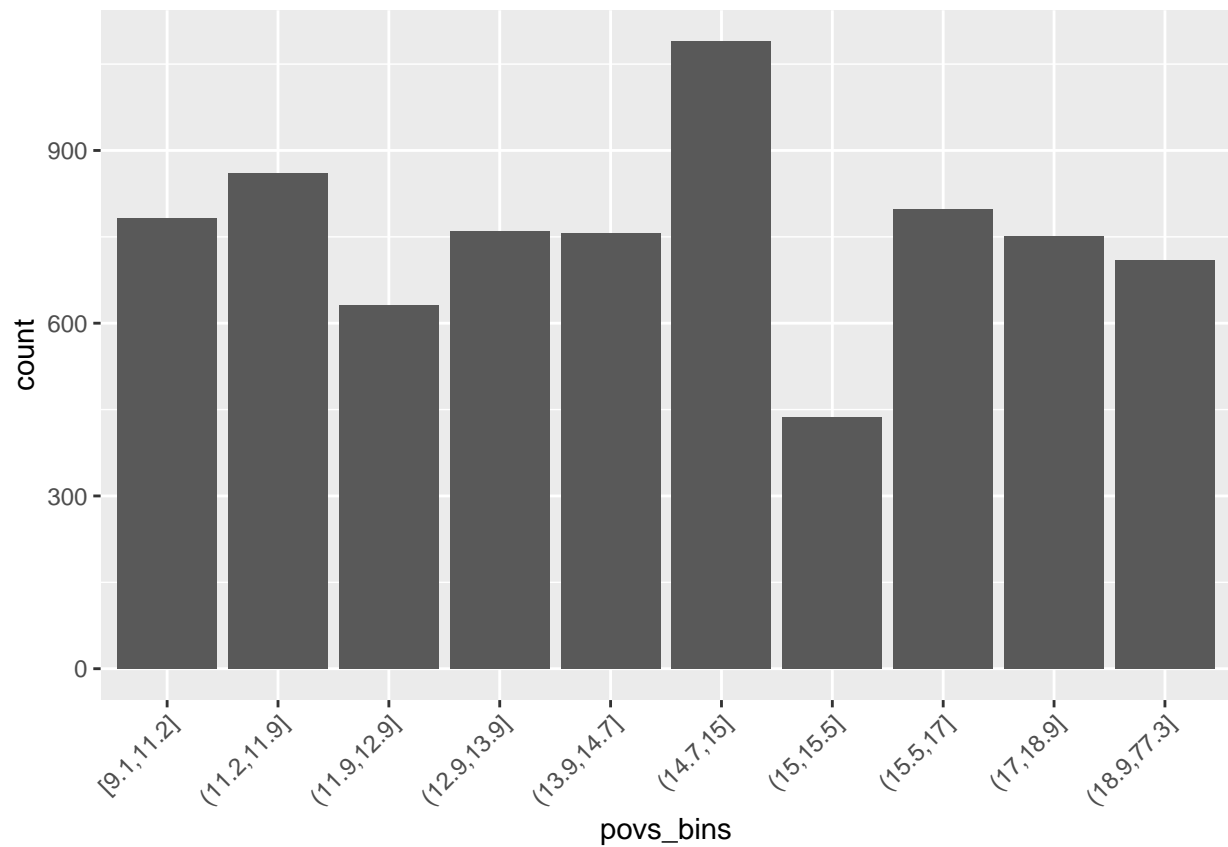
```
povs <- subset |> pull(TP6019)
subset |> ggplot(aes(x = povs)) +
  geom_histogram(binwidth = 1)
```



```
subset |> ggplot(aes(x = agemaj)) +
  geom_histogram(binwidth = 1)
```



```
povs_bins <- cut_number(povs, n = 10)
subset |> ggplot(aes(x = povs_bins)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
ages <- subset |> pull(agemaj)
ages_bins <- cut_number(ages, n = 4)
subset |> ggplot(aes(x = ages_bins)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

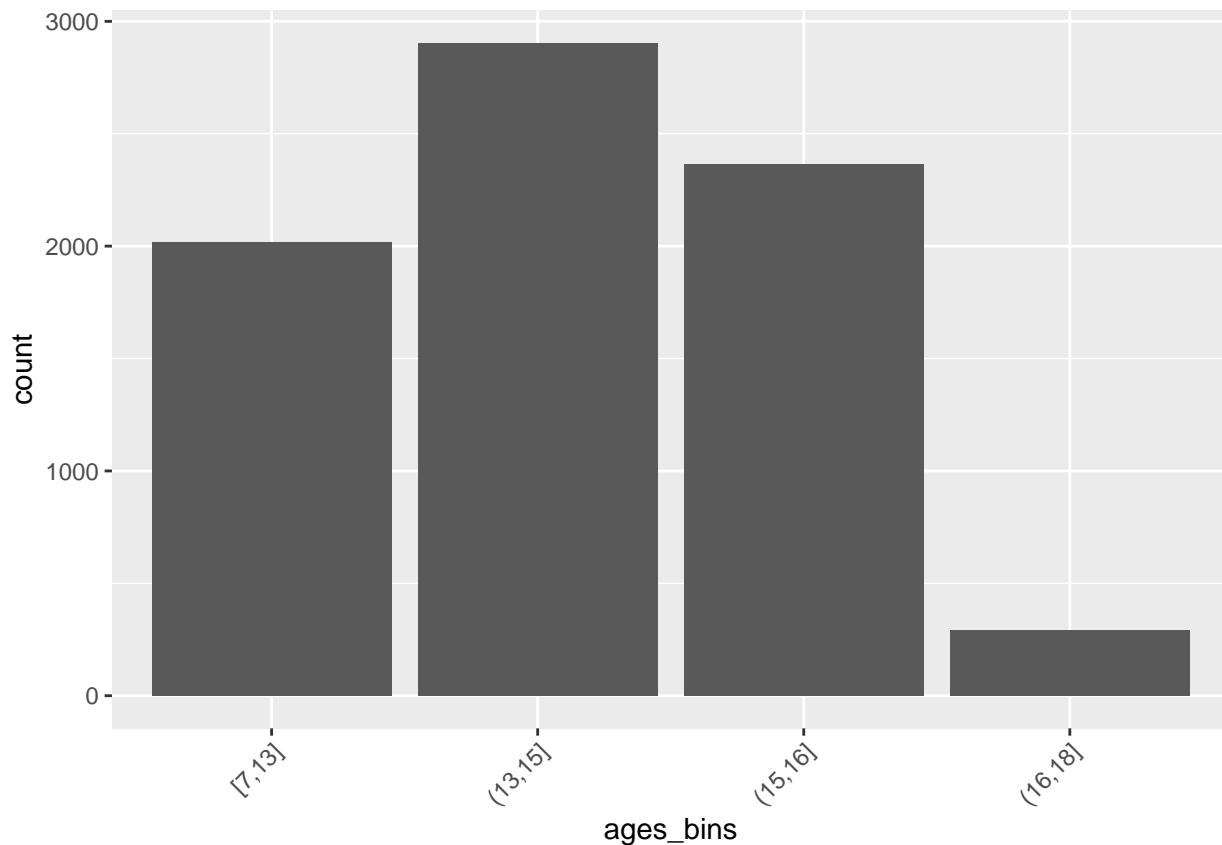


Tableau de contingence empirique.

```
table(povs_bins, ages_bins) |> addmargins()
```

```
##           ages_bins
## povs_bins  [7,13] (13,15] (15,16] (16,18] Sum
## [9.1,11.2]   171    316    259     37  783
## (11.2,11.9]  205    324    288     44  861
## (11.9,12.9]  154    238    221     19  632
## (12.9,13.9]  214    277    243     25  759
## (13.9,14.7]  214    286    228     28  756
## (14.7,15]    348    383    305     53 1089
## (15,15.5]    105    178    143     11  437
## (15.5,17]    215    315    244     24  798
## (17,18.9]    201    299    230     21  751
## (18.9,77.3]  192    288    202     28  710
## Sum         2019    2904    2363    290 7576
```

Tableau de contingence empirique normalisé par ligne (profils-lignes).

```
table(povs_bins, ages_bins) |>
  prop.table(margin = 1) |>
  addmargins() |>
  round(2)
```

```
##           ages_bins
## povs_bins  [7,13] (13,15] (15,16] (16,18] Sum
## [9.1,11.2]   0.22    0.40    0.33    0.05 1.00
## (11.2,11.9]  0.24    0.38    0.33    0.05 1.00
```

```
## (11.9,12.9] 0.24 0.38 0.35 0.03 1.00
## (12.9,13.9] 0.28 0.36 0.32 0.03 1.00
## (13.9,14.7] 0.28 0.38 0.30 0.04 1.00
## (14.7,15] 0.32 0.35 0.28 0.05 1.00
## (15,15.5] 0.24 0.41 0.33 0.03 1.00
## (15.5,17] 0.27 0.39 0.31 0.03 1.00
## (17,18.9] 0.27 0.40 0.31 0.03 1.00
## (18.9,77.3] 0.27 0.41 0.28 0.04 1.00
## Sum 2.63 3.86 3.14 0.37 10.00
```

Tableau de contingence théorique sous l'hypothèse d'indépendance.

```
ct <- chisq.test(table(povs_bins, ages_bins))
ct$expected |> round(2)
```

```
##          ages_bins
## povs_bins [7,13] (13,15] (15,16] (16,18]
## [9.1,11.2] 208.67 300.14 244.22 29.97
## (11.2,11.9] 229.46 330.03 268.55 32.96
## (11.9,12.9] 168.43 242.26 197.12 24.19
## (12.9,13.9] 202.27 290.94 236.74 29.05
## (13.9,14.7] 201.47 289.79 235.80 28.94
## (14.7,15] 290.22 417.43 339.67 41.69
## (15,15.5] 116.46 167.51 136.30 16.73
## (15.5,17] 212.67 305.89 248.90 30.55
## (17,18.9] 200.14 287.87 234.24 28.75
## (18.9,77.3] 189.21 272.15 221.45 27.18
```

La probabilité qu'une χ^2 à 27 degrés de liberté soit aussi extrême que celle observée est la suivante.

```
stopifnot(ct$p.value - (1 - pchisq(unname(ct$statistic), ct$parameter)) < 1e-10)
ct$p.value |> format(digits = 2)
```

```
## [1] "4e-04"
```

On peut donc confortablement rejeter l'hypothèse d'indépendance entre le taux de pauvreté et l'âge souhaité pour traitement comme un majeur. Ceci implique également, a fortiori, le rejet de l'hypothèse aux seuils de 10 % et 5 %.