Labo8: Statistiques bivariées

Visseho Adjiwanou, PhD.

08 March 2022

Introduction

Quand les chercheurs vont collecter les données, il ne mesure pas souvent la même variable de trois manières différentes (nominale, ordinale, ratio, intervalle). Non, ils/elles choisissent leur echelle de mesure avant d'aller sur le terrain. Aussi, pour ce labo, on a besoin de plusieurs bases de données qui ont l'information recherchée.

Nous allons travailler avec - les données de l'enquête sociale du Canada de 1995 "cora-crsc1996-E-1996_F1.csv". Vous devez lire les informations contenues dans le disctionnaire avant le cours.

Croisement de deux variables qualitatives

```
rm(list = ls())

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## Warning: package 'ggplot2' was built under R version 3.6.2
## Warning: package 'tibble' was built under R version 3.6.2
## Warning: package 'tidyr' was built under R version 3.6.2
## Warning: package 'readr' was built under R version 3.6.2
## Warning: package 'purrr' was built under R version 3.6.2
## Warning: package 'dplyr' was built under R version 3.6.2
## Warning: package 'forcats' was built under R version 3.6.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(summarytools)
```

```
## Warning: package 'summarytools' was built under R version 3.6.2
```

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

```
## Warning in system2("/usr/bin/otool", c("-L", shQuote(DSO)), stdout = TRUE):
## running command ''/usr/bin/otool' -L '/Library/Frameworks/R.framework/Resources/
## library/tcltk/libs//tcltk.so'' had status 1
```

```
## For best results, restart R session and update pander using devtools:: or remotes::install_github('ra
```

```
##
## Attaching package: 'summarytools'
```

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

```
#library(pander)

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

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double()
## )
## i Use `spec()` for the full column specifications.
```

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


crsc_classe <-
  crsc96 %>%
  select(sexq, q2)
```

Statistiques bivariées : Association entre variables

Existe-il une association entre le sexe et l'opinion des gens, notamment q2? - q2: "An unmarried girl of 18 should not have sexual relations" Une jeune fille non mariée de 18 ans ne devrait pas avoir de relations sexuelles

```
table1 <- table(crsc96_small$sexq, crsc96_small$q2)
table1
```

```
##
##       1   2   3   4   5
##   1 208 304  12 418 419
##   2 308 332  14 476 368
```

Comme vous le voyez, ce tableau n'est pas assez explicite. Il manque plusieurs éléments. On ne va pas utiliser **base R** pour l'analyse bivariée. On va utiliser le package summarytools.

Statistiques bivariées : Association entre variables

https://cran.r-project.org/web/packages/summarytools/vignettes/Introduction.html

```
table1_mieux <- ctable(crsc96_small$sexq, crsc96_small$q2, "r")
table1_mieux
```

```
## Cross-Tabulation, Row Proportions
## sexq * q2
## Data Frame: crsc96_small
##
## ------- ---- ------------ ------------ ----------- ------------ ------------ ---------------
##           q2            1            2           3            4            5           Total
##    sexq
##       1        208 (15.3%)   304 (22.3%)   12 (0.9%)   418 (30.7%)   419 (30.8%)   1361 (100.0%)
##       2        308 (20.6%)   332 (22.2%)   14 (0.9%)   476 (31.8%)   368 (24.6%)   1498 (100.0%)
##   Total        516 (18.0%)   636 (22.2%)   26 (0.9%)   894 (31.3%)   787 (27.5%)   2859 (100.0%)
## ------- ---- ------------ ------------ ----------- ------------ ------------ ---------------
```

Statistiques bivariées : Association entre variables

- Recréons la variable sexe pour qu'elle soit plus explicite.
- Recréons la question q2 pour qu'elle soit aussi plus explicite.

```
crsc96_small <-
  crsc96_small %>%
  mutate(sexe = factor(sexq, labels = c("Homme", "Femme")),
         q2_new = factor(q2, labels = c("totalement d'accord", "d'accord", "Ne sait pas", "En désaccord
```

Statistiques bivariées : Association entre variables

```
ctable(crsc96_small$sexe, crsc96_small$q2_new)
```

```
## Cross-Tabulation, Row Proportions
## sexe * q2_new
## Data Frame: crsc96_small
##
## ------- -------- -------------------- ------------ ------------ -------------- ------------------
##           q2_new   totalement d'accord      d'accord   Ne sait pas   En désaccord   Totalement en dé
##    sexe
##   Homme                    208 (15.3%)   304 (22.3%)     12 (0.9%)    418 (30.7%)              419
##   Femme                    308 (20.6%)   332 (22.2%)     14 (0.9%)    476 (31.8%)              368
##   Total                    516 (18.0%)   636 (22.2%)     26 (0.9%)    894 (31.3%)              787
## ------- -------- -------------------- ------------ ------------ -------------- ------------------
```

- Par défaut, **ctable** calcule le pourcentage ligne (row)

Statistiques bivariées : Association entre variables

Chaque commande a toujours des options. - **useNA** permet de spécifier les colonnes pour les valeurs manquantes aussi (no, ifany, always) - **round.digits** spécifie le nombre de virgule - **prop** spécifie si on calcule des proportions ligne (**r**) ou colonne (**c**) - **style** spécifie la forme du tableau (**grid, simple, rmarkdown**)

```
ctable(crsc96_small$sexe, crsc96_small$q2_new, prop = "r", style = 'rmarkdown', useNA = "no", round.digi
```

```
## ### Cross-Tabulation, Row Proportions
```

```
## #### sexe * q2_new
## **Data Frame:** crsc96_small
##
## |       |        |                     |             |             |              |
## |------:|-------:|--------------------:|------------:|------------:|-------------:|-----------------
## |       | q2_new | totalement d'accord |    d'accord | Ne sait pas | En désaccord | Totalement en dés
## |  sexe |        |                     |             |             |              |
## | Homme |        |         208 (15.3%) | 304 (22.3%) |   12 (0.9%) |  418 (30.7%) |             419
## | Femme |        |         308 (20.6%) | 332 (22.2%) |   14 (0.9%) |  476 (31.8%) |             368
## | Total |        |         516 (18.0%) | 636 (22.2%) |   26 (0.9%) |  894 (31.3%) |             787
```

Association

Les colonnes et les lignes d'un tableau croisés, ne sont pas identiques.

```
ctable(crsc96_small$q2_new, crsc96_small$sexe)
```

```
## Cross-Tabulation, Row Proportions
## q2_new * sexe
## Data Frame: crsc96_small
##
## ------------------------ ------ -------------- -------------- ---------------
##                            sexe          Homme          Femme           Total
##                  q2_new
##       totalement d'accord           208 (40.3%)    308 (59.7%)    516 (100.0%)
##                  d'accord           304 (47.8%)    332 (52.2%)    636 (100.0%)
##               Ne sait pas            12 (46.2%)     14 (53.8%)     26 (100.0%)
##             En désaccord           418 (46.8%)    476 (53.2%)    894 (100.0%)
##   Totalement en désaccord           419 (53.2%)    368 (46.8%)    787 (100.0%)
##                     Total          1361 (47.6%)   1498 (52.4%)   2859 (100.0%)
## ------------------------ ------ -------------- -------------- ---------------
```

Lequel des deux tableaux donne une indication sur l'association entre les deux variables?

Association

Aussi, est-il important de préciser si vous calculez des proportions lignes ou des proportions colonnes.

```
ctable(crsc96_small$q2_new, crsc96_small$sexe, "c")
```

```
## Cross-Tabulation, Column Proportions
## q2_new * sexe
## Data Frame: crsc96_small
##
## ------------------------ ------ --------------- --------------- ---------------
##                            sexe           Homme           Femme           Total
##                  q2_new
##       totalement d'accord           208 ( 15.3%)    308 ( 20.6%)    516 ( 18.0%)
##                  d'accord           304 ( 22.3%)    332 ( 22.2%)    636 ( 22.2%)
##               Ne sait pas            12 (  0.9%)     14 (  0.9%)     26 (  0.9%)
##             En désaccord           418 ( 30.7%)    476 ( 31.8%)    894 ( 31.3%)
##   Totalement en désaccord           419 ( 30.8%)    368 ( 24.6%)    787 ( 27.5%)
##                     Total          1361 (100.0%)   1498 (100.0%)   2859 (100.0%)
## ------------------------ ------ --------------- --------------- ---------------
```

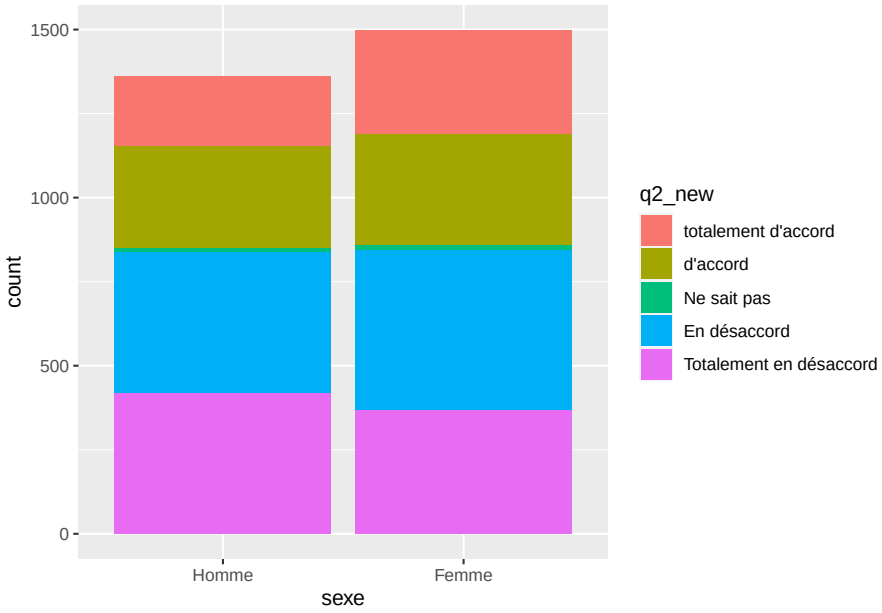
Exercices

- Créer la variable q2_3 qui regroupe les modalités de q2 en trois catégories en
  - regroupant tout ce qui est **agree** ensemble et
  - tout ce qui est **disagree** ensemble
- Regarder à nouveau l'association entre le sexe et le nouveau q2_3
- Analyser l'association entre l'age et le nouveau q2_3? Que concluez-vous?

Visualisation de l'association de deux variables qualitatives

Croisement de deux variables qualitatives

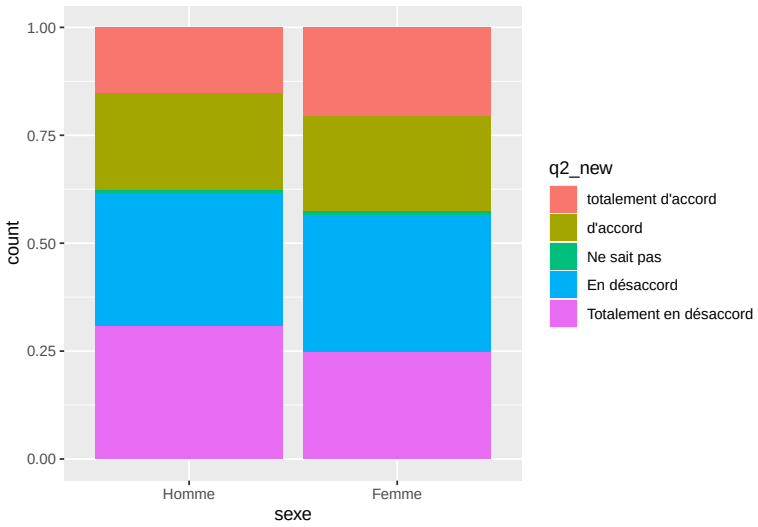
```
ggplot(crsc96_small) +
  geom_bar(aes(x = sexe, fill = q2_new))
```



- Ce graphique nous donne pour chaque sexe, le nombre de personne qui sont dans chaque catégorie de la variable dépendante.
- Il a cependant un problème, c'est difficile de comparer le nombres bruts. Il faut des pourcentages.

Croisement de deux variables qualitatives

```
ggplot(crsc96_small) +
  geom_bar(aes(x = sexe, fill = q2_new), position = "fill")
```

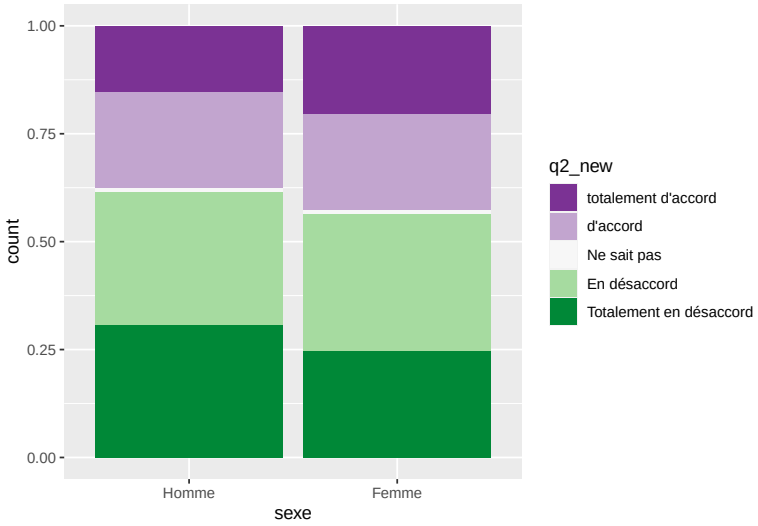


- On voit clairement la différence d'opinion entres les hommes et les femmes.

Croisement de deux variables qualitatives

- On peut changer les couleurs, on verra cela plus loin.
- http://www.sthda.com/french/wiki/couleurs-dans-r

```
ggplot(crsc96_small) +
  geom_bar(aes(x = sexe, fill = q2_new), position = "fill") +
  scale_fill_brewer(palette="PRGn")
```




```
# Changer PRGn avec un chiffre
```