Labo 6.1: Statistiques descriptives avec RStudio

Visseho Adjiwanou, PhD.

16 February 2022

## Plan

1. Fréquences, pourcentages
2. Mesures des paramètres de tendance centrale
3. Mesure des paramètres de dispersion/variation
4. Visualisation

## Données

On va travailler avec les données de l'enquête canadienne sur le revenu de 2015. <https://search1.odesi.ca/#/>

## Mesure des paramètres de dispersion

rm(list = ls())  
  
  
#library(qss)  
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.6.2

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ ggplot2 3.3.5 ✓ purrr 0.3.4  
## ✓ tibble 3.1.6 ✓ dplyr 1.0.7  
## ✓ tidyr 1.1.4 ✓ stringr 1.4.0  
## ✓ readr 1.4.0 ✓ 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)

## 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

##   
## Attaching package: 'summarytools'

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

#library(tinytex)  
  
cis\_2015 <- read\_csv("CIS-72M0003-E-2015\_F1.csv")

##   
## ── Column specification ────────────────────────────────────────────────────────  
## cols(  
## .default = col\_double(),  
## EFID = col\_character(),  
## CFID = col\_character(),  
## VERDATE = col\_character()  
## )  
## ℹ Use `spec()` for the full column specifications.

## Sélection les variables importantes et création de nouvelles variables

cis\_short <-  
 cis\_2015 %>%   
 select(PUMFID, PERSONID, AGEGP, SEX, MARST, PROV, IMMSTP, YRIMMGP, CMPHI, HLEV2G, CFID, CFSIZE, CFMJSI, CFATINC, CFEARNG, EFID, EFATINC, HHSIZE, FWEIGHT) %>%   
 mutate(Province = case\_when(  
 PROV == 10 ~ "Terre-Neuve-et-Labrador",  
 PROV == 11 ~ "Ile du prince Edouard",  
 PROV == 12 ~ "Nouvelle-Écosse",  
 PROV == 13 ~ "Nouveau-Brunswick",  
 PROV == 24 ~ "Québec",  
 PROV == 35 ~ "Ontario",  
 PROV == 46 ~ "Manitoba",  
 PROV == 47 ~ "Saskatchewan",  
 PROV == 48 ~ "Alberta",  
 PROV == 59 ~ "Colombie Britanique"  
 ),  
 region = case\_when(  
 PROV <= 13 ~ "Atlantique",  
 PROV == 24 ~ "Québec",  
 PROV == 35 ~ "Ontario",  
 PROV >=46 & PROV <= 48 ~ "Ouest",  
 PROV == 59 ~ "Colombie Britanique"  
 ),  
 sexe = if\_else(SEX == 1, "Male", "Female"))  
  
#sort(cis\_short$CFATINC)

CMPHI : Flag - Person completed high school HLEV2G : Highest level of education of person, 2nd grouping EFID : Economic family identifier EFSIZE : Number of economic family members EFATINC : EF - After-tax income HHSIZE : Number of household members FWEIGHT : Final weight

## 1. Fréquences

freq(cis\_2015$PROV)

## Frequencies   
## cis\_2015$PROV   
## Type: Numeric   
##   
## Freq % Valid % Valid Cum. % Total % Total Cum.  
## ----------- ------- --------- -------------- --------- --------------  
## 10 2255 3.76 3.76 3.76 3.76  
## 11 1527 2.54 6.30 2.54 6.30  
## 12 3260 5.43 11.73 5.43 11.73  
## 13 2972 4.95 16.68 4.95 16.68  
## 24 10840 18.06 34.74 18.06 34.74  
## 35 15823 26.36 61.10 26.36 61.10  
## 46 5764 9.60 70.70 9.60 70.70  
## 47 4908 8.18 78.88 8.18 78.88  
## 48 6340 10.56 89.44 10.56 89.44  
## 59 6339 10.56 100.00 10.56 100.00  
## <NA> 0 0.00 100.00  
## Total 60028 100.00 100.00 100.00 100.00

freq(cis\_2015$AGEGP)

## Frequencies   
## cis\_2015$AGEGP   
## Type: Numeric   
##   
## Freq % Valid % Valid Cum. % Total % Total Cum.  
## ----------- ------- --------- -------------- --------- --------------  
## 1 4038 6.73 6.73 6.73 6.73  
## 2 2774 4.62 11.35 4.62 11.35  
## 3 4191 6.98 18.33 6.98 18.33  
## 4 1429 2.38 20.71 2.38 20.71  
## 5 1107 1.84 22.55 1.84 22.55  
## 6 3296 5.49 28.05 5.49 28.05  
## 7 3314 5.52 33.57 5.52 33.57  
## 8 3654 6.09 39.65 6.09 39.65  
## 9 3761 6.27 45.92 6.27 45.92  
## 10 3699 6.16 52.08 6.16 52.08  
## 11 4059 6.76 58.84 6.76 58.84  
## 12 4663 7.77 66.61 7.77 66.61  
## 13 4834 8.05 74.66 8.05 74.66  
## 14 4353 7.25 81.92 7.25 81.92  
## 15 3800 6.33 88.25 6.33 88.25  
## 16 7056 11.75 100.00 11.75 100.00  
## <NA> 0 0.00 100.00  
## Total 60028 100.00 100.00 100.00 100.00

freq(cis\_short$Province)

## Frequencies   
## cis\_short$Province   
## Type: Character   
##   
## Freq % Valid % Valid Cum. % Total % Total Cum.  
## ----------------------------- ------- --------- -------------- --------- --------------  
## Alberta 6340 10.56 10.56 10.56 10.56  
## Colombie Britanique 6339 10.56 21.12 10.56 21.12  
## Ile du prince Edouard 1527 2.54 23.67 2.54 23.67  
## Manitoba 5764 9.60 33.27 9.60 33.27  
## Nouveau-Brunswick 2972 4.95 38.22 4.95 38.22  
## Nouvelle-Écosse 3260 5.43 43.65 5.43 43.65  
## Ontario 15823 26.36 70.01 26.36 70.01  
## Québec 10840 18.06 88.07 18.06 88.07  
## Saskatchewan 4908 8.18 96.24 8.18 96.24  
## Terre-Neuve-et-Labrador 2255 3.76 100.00 3.76 100.00  
## <NA> 0 0.00 100.00  
## Total 60028 100.00 100.00 100.00 100.00

# Paramètres de tendance centrale

## Revenu familial moyen

mean(cis\_short$CFATINC)

## [1] 80478.96

median(cis\_short$CFATINC)

## [1] 69160

mean\_income <-  
 cis\_short %>%   
 summarise(mean(CFATINC),  
 median(CFATINC))  
  
mean\_income

## # A tibble: 1 × 2  
## `mean(CFATINC)` `median(CFATINC)`  
## <dbl> <dbl>  
## 1 80479. 69160

## Revenu moyen dans chaque province

mean\_income\_province <-  
 cis\_short %>%   
 group\_by(Province) %>%   
 summarise(mean(CFATINC))  
  
mean\_income\_province

## # A tibble: 10 × 2  
## Province `mean(CFATINC)`  
## <chr> <dbl>  
## 1 Alberta 103446.  
## 2 Colombie Britanique 81159.  
## 3 Ile du prince Edouard 69475.  
## 4 Manitoba 72506.  
## 5 Nouveau-Brunswick 65471.  
## 6 Nouvelle-Écosse 67264.  
## 7 Ontario 85370.  
## 8 Québec 71151.  
## 9 Saskatchewan 86322.  
## 10 Terre-Neuve-et-Labrador 78511.

On peut aussi calculer de cette manière les autres paramètres de tendance centrale comme la médiane ...

## Revenu médian par province

md\_income\_province <-  
 cis\_short %>%   
 group\_by(Province) %>%   
 summarise(median(CFATINC))  
  
md\_income\_province

## # A tibble: 10 × 2  
## Province `median(CFATINC)`  
## <chr> <dbl>  
## 1 Alberta 85350   
## 2 Colombie Britanique 71260   
## 3 Ile du prince Edouard 62885   
## 4 Manitoba 63455   
## 5 Nouveau-Brunswick 59960   
## 6 Nouvelle-Écosse 55640   
## 7 Ontario 73850   
## 8 Québec 63045   
## 9 Saskatchewan 76758.  
## 10 Terre-Neuve-et-Labrador 66540

# Paramètres de variation

## Variance du revenu

variation\_income <-  
 cis\_short %>%   
 group\_by(Province) %>%   
 summarise(var(CFATINC, na.rm = TRUE))  
  
variation\_income

## # A tibble: 10 × 2  
## Province `var(CFATINC, na.rm = TRUE)`  
## <chr> <dbl>  
## 1 Alberta 8597415866.  
## 2 Colombie Britanique 3055789214.  
## 3 Ile du prince Edouard 1776553568.  
## 4 Manitoba 2413181777.  
## 5 Nouveau-Brunswick 1675668685.  
## 6 Nouvelle-Écosse 2144437281.  
## 7 Ontario 3566896978.  
## 8 Québec 2171717966.  
## 9 Saskatchewan 3243320891.  
## 10 Terre-Neuve-et-Labrador 3030042307.

On voit que de manière générale, la variance des revenus est aussi très grands à Alberta. On voit que les niveaux de la variance sont très grands et ne sont pas de la même unité que le revenu. C'est pourquoi, on va lui préférer l'écart-type.

ecart\_income <-  
 cis\_short %>%   
 group\_by(Province) %>%   
 summarise(sd(CFATINC, na.rm = TRUE))  
  
ecart\_income

## # A tibble: 10 × 2  
## Province `sd(CFATINC, na.rm = TRUE)`  
## <chr> <dbl>  
## 1 Alberta 92722.  
## 2 Colombie Britanique 55279.  
## 3 Ile du prince Edouard 42149.  
## 4 Manitoba 49124.  
## 5 Nouveau-Brunswick 40935.  
## 6 Nouvelle-Écosse 46308.  
## 7 Ontario 59724.  
## 8 Québec 46602.  
## 9 Saskatchewan 56950.  
## 10 Terre-Neuve-et-Labrador 55046.

Vous pouvez alors aisément calculer les autres paramètres de variation.

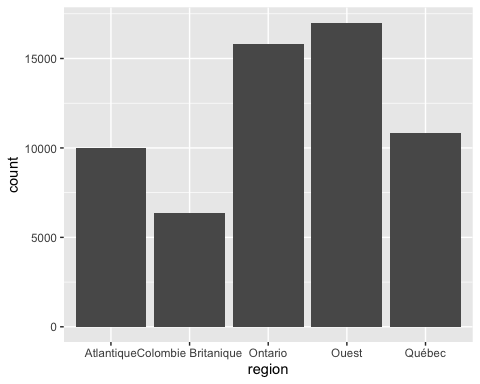
On ne peut pas comparer directement les revenus des gens de Québec avec Ontario, il faut pour cela standardisés les revenus avant de les comparer. Mais, avant de faire cela, visualisons la distribution du revenu. La visualisation est un bon moyen de se faire une première idée de la nature des données.

# Visualisation

# Variables nominales et ordinales

## Diagramme de barre

ggplot(cis\_short) +  
 geom\_bar(aes(x = region))

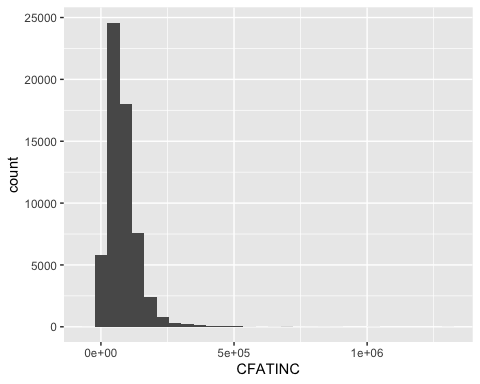


# Variable de ratio et d'intervalle

## Histogramme

ggplot(cis\_short) +  
 geom\_histogram(aes(x = CFATINC))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

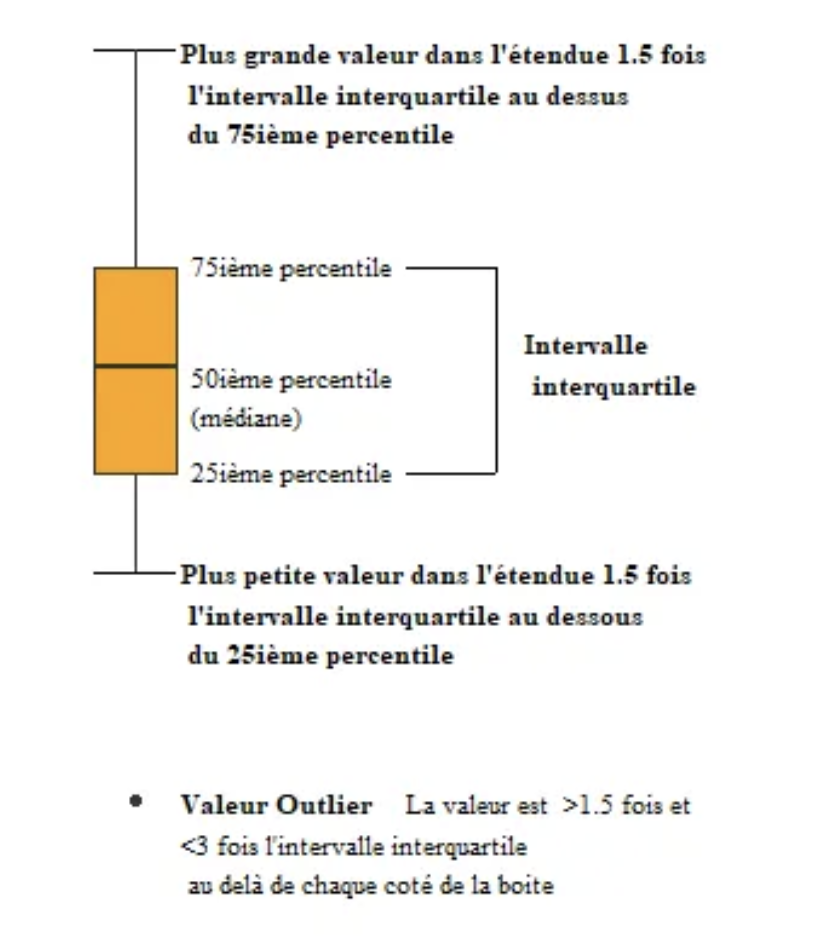


Mais, on peut présenter la distribution pour chaque province. Comment pensez-vous qu'on puisse le faire?

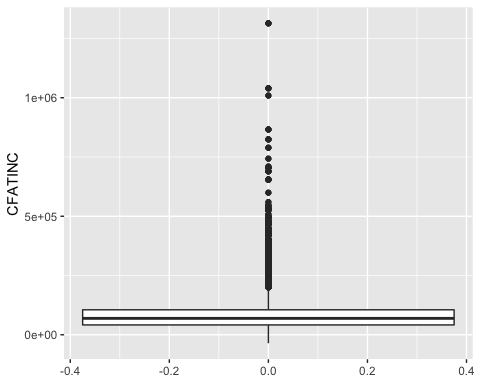
## Boxplot

<https://statistique-et-logiciel-r.com/comment-detecter-les-outliers-avec-r/>

knitr::include\_graphics("/Users/visseho/OneDrive - UQAM/Cours/Images\_cours/boxplot\_interpretation.png")

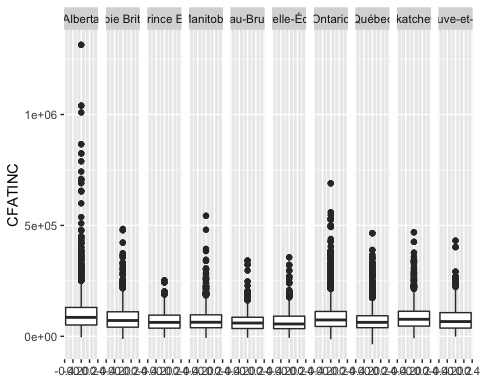


ggplot(cis\_short) +  
 geom\_boxplot(aes(y = CFATINC))



Ce graphique permet de visualiser les données abberantes ou les outliers. Un **outlier**, ou **donnée aberrante** est “une valeur ou une observation qui est « distante » des autres observations effectuées sur le même phénomène, c’est-à-dire qu’elle contraste grandement avec les valeurs « normalement » mesurées. Une donnée aberrante peut être due à la variabilité inhérente au phénomène observé ou bien elle peut aussi indiquer une erreur expérimentale. Les dernières sont parfois exclues de la série de données”. Mais avant cela, voyons comment se présentent les distributions selon les régions du Canada.

ggplot(cis\_short) +  
 geom\_boxplot(aes(y = CFATINC)) +  
 facet\_grid(~ Province)



Ajoutons-y un peu de couleurs

ggplot(cis\_short) +  
 geom\_boxplot(aes(x = Province, y = CFATINC, color = Province))

