Projet Final

Marc MESSANT, Charles ODEND'HAL

08 décembre 2023

"Nous déclarons sur l'honneur que ce mémoire a été écrit de notre main, sans aide extérieure non autorisée, qu'il n'a pas été présenté auparavant pour évaluation et qu'il n'a jamais été publié, dans sa totalité ou en partie.

Toutes parties, groupes de mots ou idées, aussi limités soient-ils, y compris des tableaux, graphiques, cartes etc. qui sont empruntés ou qui font référence à d'autres sources bibliographiques sont présentés comme tels, sans exception aucune."

Signature
Marc MESSANT, Charles ODEND'HAL

Contents

Introduction

Le vol de véhicules est une problématique sociale majeure en France, ayant des implications directes sur la sécurité des citoyens et suscitant des préoccupations croissantes. Comprendre les facteurs qui influent sur la fréquence de ces délits revêt une importance cruciale pour le développement de stratégies de prévention efficaces et la mise en place d'une sécurité publique adaptée.

L'analyse territoriale constitue un axe central de notre démarche, visant à identifier les tendances départementales des vols de véhicules. Cette approche permettra d'obtenir des informations spécifiques à chaque département, offrant ainsi une vision détaillée de la situation géographique de ce délit.

1. Présentation du projet

Notre projet d'étude se propose d'explorer la corrélation entre les vols de véhicules répertoriés dans les différents départements de France et plusieurs variables socio-économiques, comme notamment le nombre de policiers dans chacun de ces départements ou encore le revenu moyen d'un habitant vivant dans ce département.

L'inclusion du nombre de policiers par département dans notre analyse vise à déterminer l'impact des effectifs policiers sur la fréquence des vols de véhicules. Ces résultats pourraient fournir des indications précieuses quant à l'efficacité des forces de l'ordre et orienter les décideurs vers des ajustements nécessaires dans la répartition des ressources en matière de sécurité publique. On peut imaginer qu'un nombre important de policiers signifierait un nombre de vols de véhicules plus faible.

Notre projet tire ses sources de quatre bases de données

Base statistique départementale de la délinquance enregistrée par la police et la gendarmerie nationale entre 2016 et 2022: https://www.data.gouv.fr/fr/datasets/bases-statistiques-communale-et-departementale-de-la-delinquance-enregistree-par-la-police-et-la-gendarmerie-nationales/

Nombre départemental de policiers et gendarmes pour 10 000 habitants en 2019: https://www.insee.fr/fr/statistiques/5763601?sommaire=5763633

Taux de chômage départemental par trimestre entre 2016 et 2022: https://www.insee.fr/fr/statistiques/2012804

Revenu salarial net annuel départemental moyen en 2019: https://www.insee.fr/fr/statistiques/2012748

A partir de ces différentes bases, nous avons extrait les données pertinentes pour notr étude (une feuille de chacun des quatre fichiers) que nous avons regroupé en un seul fichier Excel nommé Base.xslx.

Le traitement initial de cette base nous a permis de pouvoir manipuler les données lors de l'étude.

2. Description des données

La population est composée de l'ensemble des départements de France, de 2016 à 2022. On compte 100 départements (hors Mayotte, car nous ne disposions pas de suffisament de données). Les bases de données accessibles sur internet ne nous ont permis de récolter les données de revenu moyen et nombre de policiers sur l'année 2019 uniquement. Les observations de notre étude seront réalisées en conséquence.

Les variables présentes dans notre étude :

- Nombre de vols de véhicules pour mille habitants : variable quantitative continue
- Nombre de policiers pour mille habitants (uniquement en 2019) : variable quantitative discrete
- Revenu moyen d'un habitant (uniquement en 2019) : variable quantitative continue
- Taux de chômage : variable quantitative discrete
- Taux de chômage regroupé en classe : variable qualitative discrete

Nettoyage de l'environnement

```
rm(list=ls())
```

Réglage du nombre de décimales

```
options(digits=3)
```

Chargement des différentes bases

```
setwd("D:/MARCO/COURS/DAUPHINE/L3/S1/Statistiques/Projet final")
library(readxl)
```

-> Traitement de la base de l'ensemble des délits, pour garder uniquement les vols de véhicules (suppression de la dernière ligne : Mayotte 976)

```
Vols <- read_excel("MESSANT_ODENDHAL_base.xlsx",sheet="Vols de véhicules",
                   range="A5658:K6364",col_names = FALSE)
Vols <- as.data.frame(Vols)
Vols <- subset(Vols, select = -c(1,4,5,6,7))
colnames(Vols) <- c("Année", "Code département", "Faits","POP", "LOG",
                     "Taux pour mille habitants")
Vols <- subset(Vols, `Code département` != 976)
```

-> Création d'une sous base du nombre de vol de véhicules en 2019

```
Vols19<-subset(Vols,Année=="19")
```

-> Traitement de la base des revenus moyens par habitant (suppression de lignes qui ne nous intéressent pas dans cette étude : France métropolitaine hors Île-de-France - 97, France métropolitaine - 98 et France hors Mayotte - 103)

```
Revenus <- read_excel("MESSANT_ODENDHAL_base.xlsx", sheet ="Revenus",
                      range="A5:C107", col_names = FALSE)
Revenus <- as.data.frame(Revenus)
Revenus <- Revenus[-c(97, 98, 103), ]
colnames(Revenus) <- c("Code département", "Département",
                       "Revenu moyen en euros")
```

-> Traitement de la base du nombre de policiers (suppression de la dernière ligne : Mayotte 976)

```
Policiers <- read_excel("MESSANT_ODENDHAL_base.xlsx", sheet="Policiers",
                        range="A3:C103", col_names = FALSE)
Policiers <- as.data.frame(Policiers)
Policiers <- Policiers[-nrow(Policiers), ]
colnames(Policiers) <- c("Code département", "Département",
                     "Nombre de policiers pour 10000 habitants")
```

-> Traitement de la base du taux de chômage : moyenne des trimestres pour avoir des données annuelles, mise en forme de toutes les années sur une colonne

```
Chomage <- read_excel("MESSANT_ODENDHAL_base.xlsx", sheet = "Taux chômage",
                      range="A5:FJ104", col_names = FALSE)
Chomage <- as.data.frame(Chomage)
Chomage <- subset(Chomage, select = -c(3:138))
mat <- matrix(nrow = 100, ncol = 28)
```

```r
mat <- subset(Chomage, select = -c(1,2))
nouv_mat<- matrix(nrow = 100, ncol = 7)
for (i in 1:100) {
  k=1
  for (j in 1:7) {
    nouv_mat[i,j] <- (mat[i,k]+mat[i,k+1]+mat[i,k+2]+mat[i,k+3])/4
    k=k+4
  }}
Chomage <- subset(Chomage, select = c(1,2))
Chomage <- cbind(Chomage,nouv_mat)
colnames(Chomage) <- c("Code département", "Département","2016","2017",
                       "2018","2019","2020","2021", "2022")
library(tidyr)
```

```
Warning: le package 'tidyr' a été compilé avec la version R 4.3.2
```

```r
Chomage <- gather(Chomage, key = "Année", value = "Taux de chômage", -c(1, 2))
```

Pour chaque variable de notre étude, nous réalisons:

- Tableau de fréquence
- Tableau de proportion
- Graphique de distribution empirique

2.1. Variable "Taux de vols de véhicules pour mille habitants" : Tableaux et graphique

```r
TxVols_bornes<-seq(0.2,6.2,0.6)
TauxVols<-Vols$`Taux pour mille habitants`
Volscut<-cut(TauxVols,TxVols_bornes)
tab <- table(Volscut)
knitr::kable(t(tab),"simple",
      caption="Tableau de fréquence du taux pour mille habitants (en classe)",
      align="c")
```

Table 1: Tableau de fréquence du taux pour mille habitants (en classe)

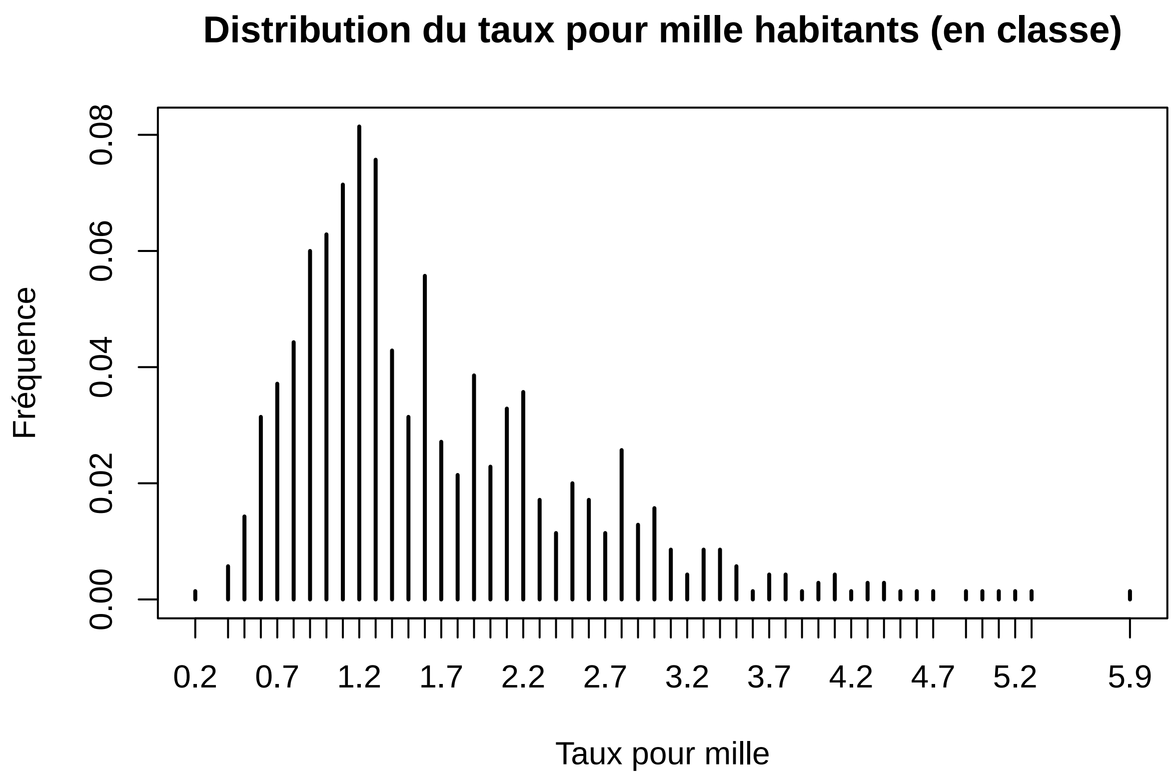
(0.2,0.8]	(0.8,1.4]	(1.4,2]	(2,2.6]	(2.6,3.2]	(3.2,3.8]	(3.8,4.4]	(4.4,5]	(5,5.6]	(5.6,6.2]
75	280	144	94	62	25	9	7	3	1

```r
tab <- prop.table(table(Volscut))
knitr::kable(t(tab),
      caption="Tableau de proportion du taux pour mille habitants (en classe)",
      align= "c")
```

Table 2: Tableau de proportion du taux pour mille habitants (en classe)

(0.2,0.8]	(0.8,1.4]	(1.4,2]	(2,2.6]	(2.6,3.2]	(3.2,3.8]	(3.8,4.4]	(4.4,5]	(5,5.6]	(5.6,6.2]
0.107	0.4	0.206	0.134	0.089	0.036	0.013	0.01	0.004	0.001

```
txVols<-round(Vols$"Taux pour mille habitants",digits=1)
plot(prop.table(table(txVols)),
     main="Distribution du taux pour mille habitants (en classe)",
     xlab="Taux pour mille", ylab="Fréquence")
```



2.2. Variable "Taux de chômage" : Tableau et graphique

```
TxChom_seq<-seq(4,24,3)
TxChom_cut<-cut(Chomage$`Taux de chômage`,TxChom_seq)
tab <- table(TxChom_cut)
knitr::kable(t(tab),"simple", caption="Tableau de fréquence du taux de chômage",
             align="c")
```

Table 3: Tableau de fréquence du taux de chômage

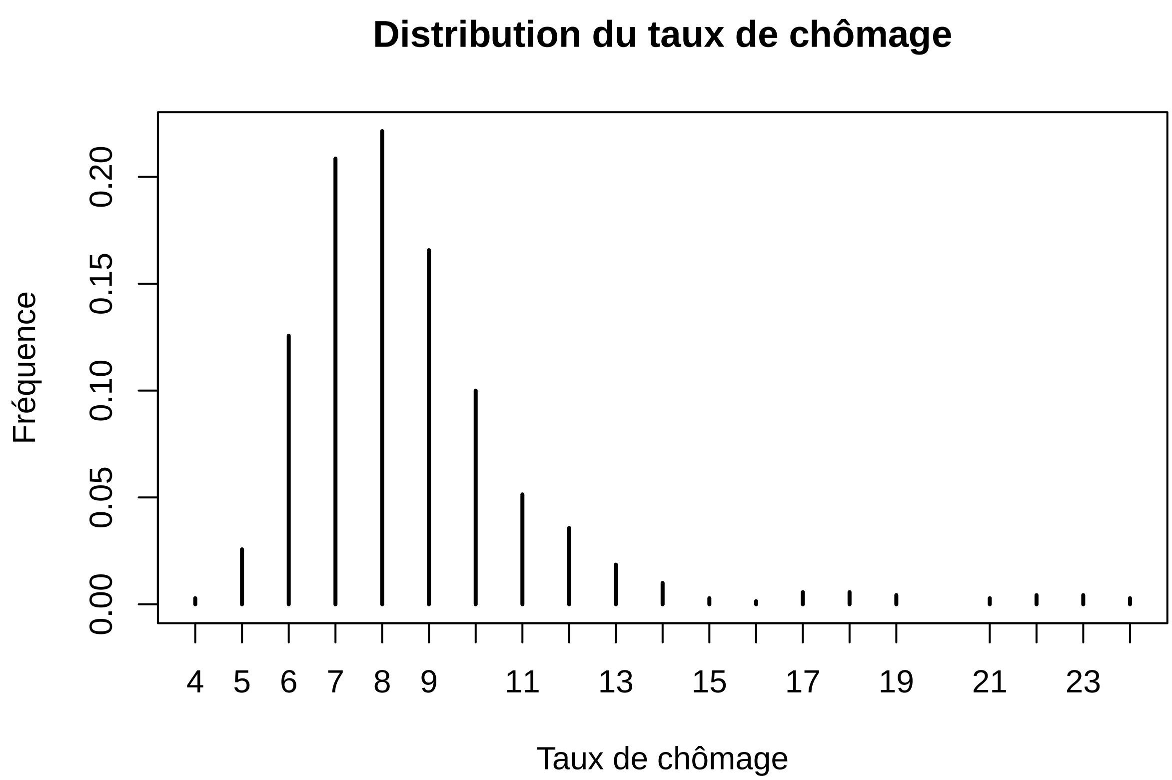
(4,7]	(7,10]	(10,13]	(13,16]	(16,19]	(19,22]
175	391	101	11	11	4

```r
tab <- prop.table(table(TxChom_cut))
knitr::kable(t(tab), caption="Tableau de proportion du taux de chômage",
             align= "c")
```

Table 4: Tableau de proportion du taux de chômage

(4,7]	(7,10]	(10,13]	(13,16]	(16,19]	(19,22]
0.253	0.564	0.146	0.016	0.016	0.006

```r
txChom<-round(Chomage$"Taux de chômage",digits=0)
plot(prop.table(table(txChom)), main="Distribution du taux de chômage",
     xlab="Taux de chômage", ylab="Fréquence")
```



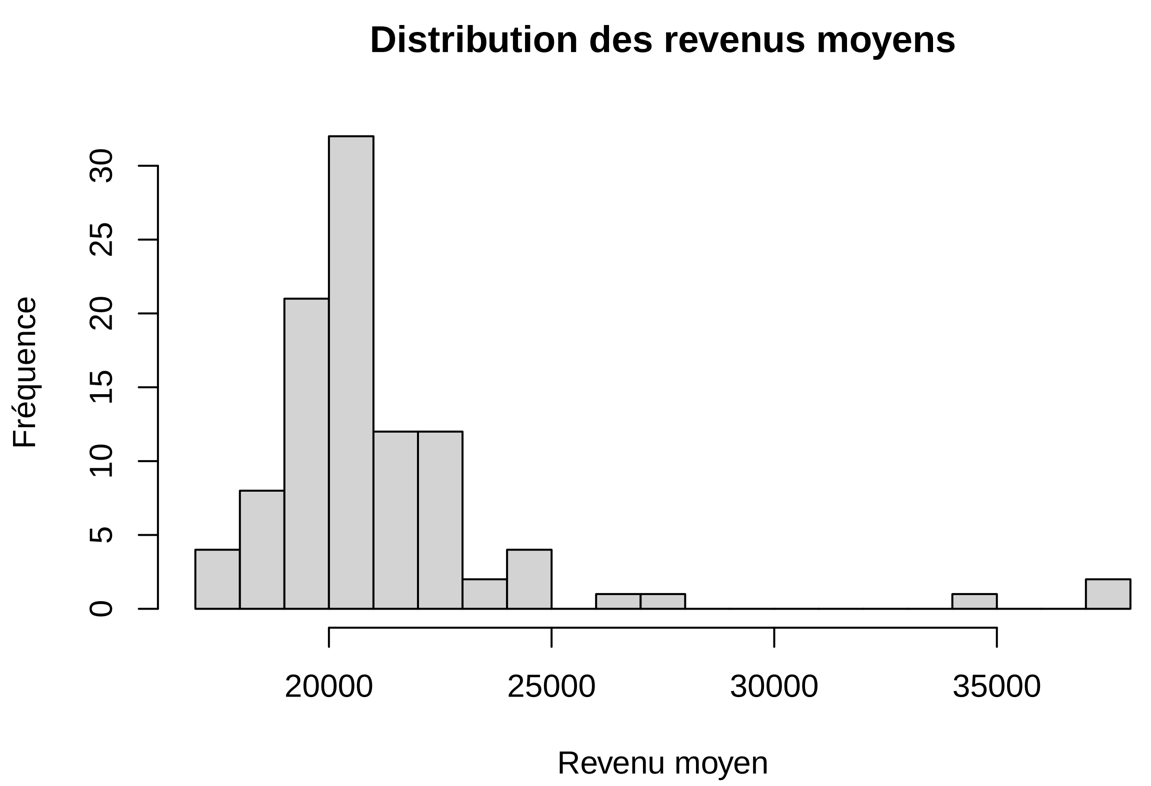
2.3. Variable "Revenus moyens" : Tableau et graphique

```
limites_classes <- c(10000, 20000, 30000, 40000)
classes <- cut(Revenus$`Revenu moyen en euros`, breaks = limites_classes,
               include.lowest = TRUE)
table_classes <- table(classes)
knitr::kable(t(table_classes),"simple",
             caption="Tableau de fréquence des revenus moyens", align="c")
```

Table 5: Tableau de fréquence des revenus moyens

[1e+04,2e+04]	(2e+04,3e+04]	(3e+04,4e+04]
33	64	3

```
hist(Revenus$`Revenu moyen en euros`, main='Distribution des revenus moyens',
     xlab="Revenu moyen", ylab="Fréquence", breaks = c(20))
```



2.4. Variable "Nombre de policiers" : Tableau et graphique

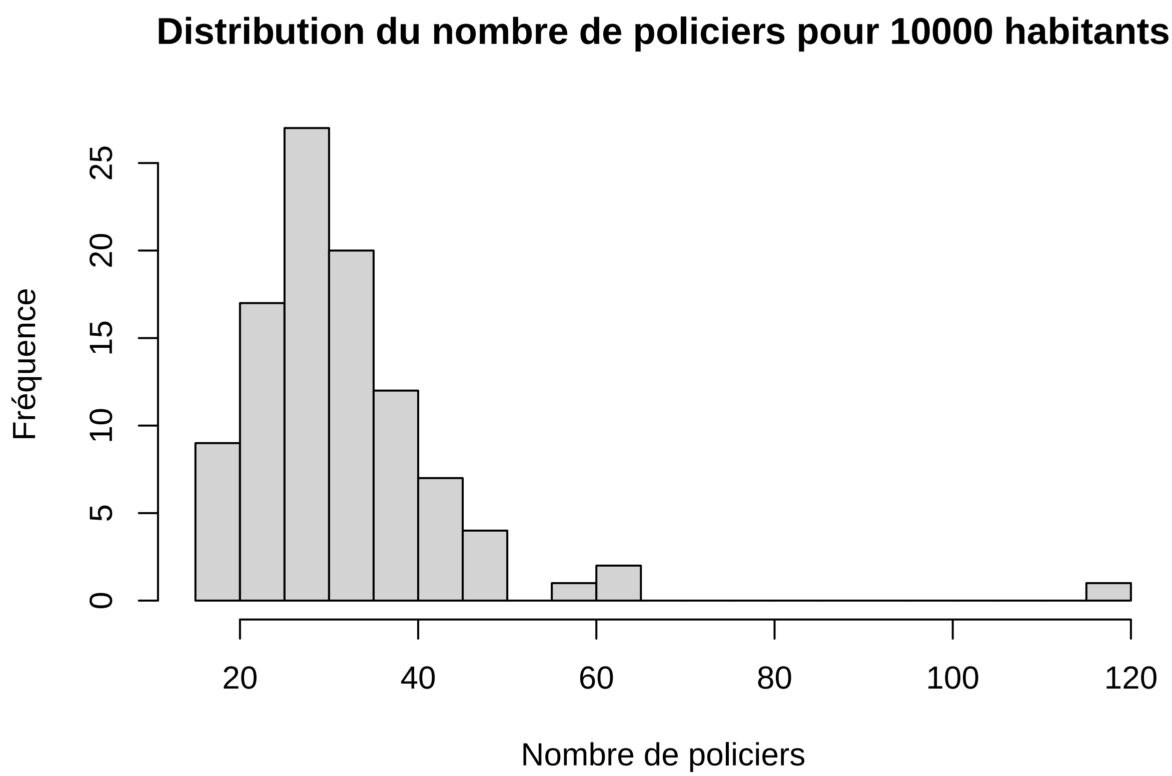
```
limites_classes <- c(10, 25, 40, 55, 70, 120)
classes <- cut(Policiers$`Nombre de policiers pour 10000 habitants`,
               breaks = limites_classes, include.lowest = TRUE)
```

```r
table_classes <- table(classes)
knitr::kable(t(table_classes),"simple",
             caption="Tableau de fréquence des revenus moyens", align="c")
```

Table 6: Tableau de fréquence des revenus moyens

[10,25]	(25,40]	(40,55]	(55,70]	(70,120]
26	59	11	3	1

```r
hist(Policiers$`Nombre de policiers pour 10000 habitants`,
     main='Distribution du nombre de policiers pour 10000 habitants',
     xlab="Nombre de policiers", ylab="Fréquence",breaks = c(20))
```



3. Estimation ponctuelle et par intervalle de confiance

Nous commençons par tirer un échantillon de 200 individus au sein de notre base. Nous assignons à la variable échantillon_vols les taux de vols associés aux valeurs de l'échantillon.

3.1. Tirage d'un écha²ntillon de 200 individus dans la population

```
n <- 200
N<-nrow(Vols)
tirages <- sample(1:N, n, replace=TRUE)
echantillon_vols <- Vols$`Taux pour mille habitants`[tirages]
```

3.2.1. Estimation ponctuelle de la moyenne empirique du taux départemental de vols de véhicules pour 1000 logements entre 2016 et 2022

```
moyenne_empirique<-mean(echantillon_vols)
```

La moyenne empirique obtenue par notre échantillon est de 1.662.

3.2.2. Estimation ponctuelle de la variance empirique corrigée du taux départemental de vols de véhicules pour 1000 logements entre 2016 et 2022

```
variance_empirique_corrigée<-var(echantillon_vols)
```

3.3. Estimation par intervalle de confiance à 90% de la moyenne théorique

Dans ce cas n=200, donc n>30, et la variance théorique n'est pas connue. Nous utilisons donc les quantiles de la loi normale pour réaliser notre intervalle.

```
borne_inf90<-moyenne_empirique-qnorm(0.95)*sd(echantillon_vols)/sqrt(n)
borne_sup90<-moyenne_empirique+qnorm(0.95)*sd(echantillon_vols)/sqrt(n)

print(c("Un intervalle de confiance à 90% est [", borne_inf90, " ; ",
        borne_sup90, "]"),
       collapse = "")
```

```
[1] "Un intervalle de confiance à 90% est ["
[2] "1.54314120692991"
[3] " ; "
[4] "1.75955330429409"
[5] "]"
```

On peut dire que nous sommes confiants à 90% que la moyenne théorique de la population se trouve dans cette intervalle.

3.4. Estimation par intervalle de confiance à 95% de la moyenne théorique du taux de vols par départements

```
borne_inf95<-moyenne_empirique-qnorm(0.975)*sd(echantillon_vols)/sqrt(n)
borne_sup95<-moyenne_empirique+qnorm(0.975)*sd(echantillon_vols)/sqrt(n)

print(c("Un intervalle de confiance à 95% est [", borne_inf95, " ; ",
        borne_sup95, "]"),
       collapse = "")
```

```
[1] "Un intervalle de confiance à 95% est ["
[2] "1.52241179603597"
[3] " ; "
[4] "1.78028271518803"
[5] "]"
```

En augmentant le niveau de confiance à 95%, nous obtenons un intervalle plus large. Nous sommes davantage confiants (ici à 95%) que la moyenne théorique de la population se trouve dans cet intervalle.

4. Test de conformité

D'après nos observations sur la moyenne empirique, nous décidons de réaliser un test de conformité sur l'échantillon plus large dont nous disposons. Nous effectuons un test unilatéral à droite (écart-type inconnu) avec comme statistique de test la moyenne empirique du taux de vols départemental pour 1000 habitants sur toute la base de données dont nous disposons, soit 700 individus.

Nous souhaitons tester la valeur de la moyenne empirique obtenue précédemment lors de l'estimation ponctuelle. Celle_ci était de environ 1.662.

Sous H0, la valeur de la moyenne est donc de 1.662.

$$\begin{cases} H_0 : m = 1.662 \\ H_1 : m \neq 1.662 \end{cases}$$

```r
x<-Vols$"Taux pour mille habitants"
t.test(x, mu = 1.662, alternative = "two.sided")
```

```
    One Sample t-test

data:  x
t = 0.04, df = 699, p-value = 1
alternative hypothesis: true mean is not equal to 1.66
95 percent confidence interval:
 1.60 1.73
sample estimates:
mean of x
     1.66
```

La valeur de la probabilité critique du test est de 1. Cette valeur est supérieure aux seuil usuels de 5% et 10%. Au vu de ce résultat, on ne rejette pas H0 aux seuils de 5 et 10%.
D'après ce test de conformité, l'estimation de la moyenne empirique de 1.662 est très bonne.

5. Test de comparaison

Nous cherchons à étudier la relation entre le nombre de policiers par départements et le taux de vols véhicules départemental sur l'année 2019.

5.1. Obtention de 2 séries de données

Nous commençons par créer une nouvelle base (BaseA). Celle-ci équivaut aux vols de 2019 (Vols19), comportant notamment les taux de vols de l'année 2019, à laquelle on rajoute le nombre de policiers par département, également sur l'année 2019.

```
BaseA<-cbind(Vols19, "Nombre de policiers pour 10000 habitants"
             =Policiers$`Nombre de policiers pour 10000 habitants`)
```

Nous créons ensuite 2 différentes bases pour isoler les séries de données. La BaseB comporte les données de la BaseA ordonnées en fonction de leur nombre de policiers dans l'ordre décroissant.

```
BaseB<-BaseA[order(-BaseA$`Taux pour mille habitants`),]
```

On reporte ensuite dans la variable top30 les valeurs des taux de vols correspondant aux 30 départements dont le nombre de policiers pour 10000 habitants est le plus élevé.

```
top30 <-BaseB[1:30,7]
```

Réciproquement, nous créons BaseC, qui classe les lignes de BaseA en fonction de leur nombre de policiers par 10000 habitants, dans l'ordre croissant.

```
BaseC<-BaseA[order(BaseA$`Taux pour mille habitants`),]
```

On reporte dans la valeur bottom30 les valeurs des taux de vols correspondant aux 30 départements dont le nombres de policiers pour 10000 habitants est le plus faible.

```
bottom30 <-BaseC[1:30,7]
```

Nous réalisons un test de variances sur les 2 séries.

```
var.test(top30,bottom30,alternative="two.sided")
```

```
    F test to compare two variances

data:  top30 and bottom30
F = 3, num df = 29, denom df = 29, p-value = 0.001
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.63 7.18
sample estimates:
ratio of variances
              3.42
```

La probabilité critique du test de variances est égale 0.001. On rejette l'hypothèse d'égalité des variances. On réalise donc un test de comparaison unilatéral, avec n=30 et var.equal=FALSE car les variances sont différentes. On cherche à déterminer si les moyennes des séries top30 et bottom30 sont égales ou non.

$$\begin{cases} H_0 : m_1 = m_2 \\ H_1 : m_1 \neq m_2 \end{cases}$$

```r
t.test(top30, bottom30, alternative ="greater",var.equal=FALSE)
```

```
    Welch Two Sample t-test

data:  top30 and bottom30
t = 2, df = 45, p-value = 0.03
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.793   Inf
sample estimates:
mean of x mean of y
     36.6      29.8
```

La probabilité critique du test est de 0.03. Cette valeur est inférieure aux seuils usuels de 5% et 10%. On rejette l'hypothèse nulle H0 aux seuils de 5% et 10% et on conclut qu'il existe une différence de taux de vols entre les départements avec le plus de policiers et les départements avec le moins de policiers par 10000 habitants en France en 2019.

6. Test du Chi-Deux

6.1. Définition des variables

Nous commençons par définir les variables TauxVols et TauxChômage.

```
TauxVols<-Vols$`Taux pour mille habitants`
TauxChômage<-Chomage$`Taux de chômage`
```

6.2. Discrétisation des variables

Ces variables sont continues. Afin de pouvoir réaliser notre test du Chi-Deux, nous devons discrétiser ces variables pour obtenir des classes dont tous les effectifs sont supérieurs à 4.

```
TxVols_bornes<-seq(0.2,6,1)
TxChômage_bornes<-seq(4,24,4)
TxVols_cut<-cut(TauxVols,TxVols_bornes)
TxChômage_cut<-cut(TauxChômage,TxChômage_bornes)
```

6.3. Tableau de contingence empirique

Nous réalisons le tableau de contingence empirique correspondant à ces classes.

```
TC_emp<-table(TxChômage_cut,TxVols_cut)
addmargins(TC_emp)
```

```
             TxVols_cut
TxChômage_cut (0.2,1.2] (1.2,2.2] (2.2,3.2] (3.2,4.2] (4.2,5.2] Sum
      (4,8]         180       103        42        11         1 337
      (8,12]         76       156        48        19         9 308
      (12,16]         0        16        13         1         1  31
      (16,20]         3         6         3         0         0  12
      (20,24]         0         6         3         1         0  10
      Sum           259       287       109        32        11 698
```

Nous obtenons le tableau suivant:

Afin de pouvoir réaliser un test du Chi-Deux, les effectifs empiriques de chaque classe doivent être supérieurs à 4. Après avoir réalisé un premier test du chi2 et testé avec la ligne de code chi2$expected, nous observons que ce n'est pas encore le cas. Nous devons donc procéder à des regroupements de classes.

6.4. Regroupements de classes

Nous commençons par regrouper les 4 classes avec des taux de vols supérieurs à 2,2 par 1000 habitants en une seule classe, nommée (0.2,1.2].

```
TxVols_cut2 <- factor(TxVols_cut,
                      levels = c("(0.2,1.2]", "(1.2,2.2]", "(2.2,3.2]",
                                 "(3.2,4.2]", "(4.2,5.2]", "(5.2,6.2]"),
                      labels = c("(0.2,1.2]", "(1.2,2.2]", "(2.2,6.2]",
                                 "(2.2,6.2]", "(2.2,6.2]", "(2.2,6.2]"))
```

Ensuite, nous regroupons les 3 classes avec un taux de chômage supérieur à 12% en une seule.

```
TxChômage_cut2 <- factor(TxChômage_cut,
                         levels = c("(4,8]", "(8,12]", "(12,16]", "(16,20]",
                                    "(20,24]"),
                         labels = c("-8%", "8-12%", "+12%", "+12%","+12%"))
```

6.5. Tableau de profils-lignes

Nous établissons un tableau de profils-lignes.

```
TC_emp2<-table(TxChômage_cut2,TxVols_cut2)
TPL<-addmargins(prop.table(addmargins(TC_emp2,1,FUN=sum),1),2,FUN=sum)
```

6.6. Tableau de profils-colonnes

Nous établissons un tableau de profils-colonnes.

```
TPC<-addmargins(prop.table(addmargins(TC_emp2,2,FUN=sum),2),1,FUN=sum)
```

6.7. Test d'indépendance du Chi-Deux

Ensuite, nous effectuons un test d'indépendance du Chi-Deux portant sur le taux de vols et le taux de chômage. On ne fait pas de correction de continuité.

```
chi2bis<-chisq.test(TxVols_cut2,TxChômage_cut2,correct=FALSE)
```

On lit une probabilité critique très faible. On rejette donc l'hypothèse d'indépendance aux seuils de 5% et 10%. Cela confirme notre intuition intiale étant qu'il existe une relation entre les vols de véhicules et le taux de chômage.

D'après ce test chi2bis$expected, nos effectifs joints sont bien tous supérieurs à 4, ce qui nous permet d'utiliser la loi du Chi-Deux dans le test.

Enfin, nous réalisons une dernière observation portant sur les coefficients de corrélation liant le taux de véhicules départemental à 3 de nos principales variables: le taux de chômage, le revenu moyen et le nombre de policiers par 10000 habitants de chaque département

6.8. Coefficients de corrélation

```
cor(TauxVols,TauxChômage)
```

```
[1] 0.275
```

```
Revenu2<-Revenus$`Revenu moyen en euros`
vols19<-Vols19$`Taux pour mille habitants`
cor(vols19,Revenu2)
```

```
[1] 0.487
```

```r
cor(vols19,Policiers$`Nombre de policiers pour 10000 habitants`)
```

```
[1] 0.327
```

Nous obtenons 3 coefficients de corrélation positifs mais relativement faibles. Nous concluons en disant qu'il existe une corrélation relativement entre chacune de ces 3 variables et le taux départemental de vols de véhicules pour 1000 habitants.

Conclusion

Pour conclure notre étude, il est difficile d'affirmer une corrélation claire entre les différentes variables.

A travers cette étude, nous avons montré qu'il existe à l'échelle départementale un lien entre le taux de vols de véhicules et le revenu moyen, le nombre de policiers par habitant et le taux de chômage.

Ces observations sont venues confirmer certaines de nos hypothèses initiales. La faible corrélation positive reliant taux de chômage et taux de vols peut s'expliquer par une plus forte criminalité dans les départements les plus touchés par le chômage. Dans les départements où le revenu moyen est important, on imagine que les opportunités de vols de véhicules sont plus fréquentes ou plus intéressantes, voire que les malfaiteurs ont davantage de moyens pour les éxécuter.

Néanmoins, notre étude a également contredit une de nos suppositions. Nous avons été surpris par la corrélation positive entre taux de vols et nombre de policiers. Celle-ci s'explique peut-être par la nécessité d'une plus grande concentration de forces de l'ordre dans les départements où la criminalité est la plus importante.