

Rapport - Statistiques

ESPANA GUTIERREZ Pablo, MULOT Lendy

08/03/2020

Introduction

Nous avons choisi de nous intéresser aux données du gouvernement français sur les émissions de polluants des véhicules commercialisés en France en 2014 (cf. Bibliographie). Nous avons décidé pour mener nos expériences, de ne pas considérer les voitures de la marque Mercedes car elles étaient surreprésentées (87% des voitures de notre jeu de données) et pourraient donc introduire un biais.

```
t2014 <- read.csv("2014.csv", header=T, sep=";", dec=",")
```

```
t <- t2014 %>%  
  filter(lib_mrq != "MERCEDES")
```

```
head(t)
```

```
##      lib_mrq lib_mod_doss lib_mod      dscom      cnit  
## 1 ALFA-ROMEO      159      159 159 1750 Tbi (200ch) M10ALFVP000G340  
## 2 ALFA-ROMEO      159      159 159 1750 Tbi (200ch) M10ALFVP000H341  
## 3 ALFA-ROMEO      159      159 159 2.0 JTDm (136ch) M10ALFVP000E302  
## 4 ALFA-ROMEO      159      159 159 2.0 JTDm (136ch) M10ALFVP000F303  
## 5 ALFA-ROMEO      159      159 159 2.0 JTDm (170ch) M10ALFVP000G304  
## 6 ALFA-ROMEO      159      159 159 2.0 JTDm (170ch) M10ALFVP000H305  
##      tvv cod_cbr hybride puiss_admin_98 puiss_max typ_boite_nb_rapp  
## 1 939AXN1B52C      ES      non      12      147      M 6  
## 2 939BXN1B53C      ES      non      12      147      M 6  
## 3 939AXR1B64      GO      non      7      100      M 6  
## 4 939AXR1B64B      GO      non      7      100      M 6  
## 5 939AXS1B66      GO      non      9      125      M 6  
## 6 939AXS1B66B      GO      non      9      125      M 6  
##      conso_urb conso_exurb conso_mixte co2 co_typ_1      hc      nox hcnox ptcl  
## 1      11.3      5.8      7.8 182      0.647 0.052 0.032      NA 0.002  
## 2      11.5      6.0      8.0 186      0.647 0.052 0.032      NA 0.002  
## 3      6.6      4.2      5.1 134      0.066      NA 0.149 0.175 0.001  
## 4      6.6      4.2      5.1 134      0.066      NA 0.149 0.175 0.001  
## 5      6.9      4.3      5.3 139      0.060      NA 0.164 0.193 0.001  
## 6      6.9      4.3      5.3 139      0.060      NA 0.164 0.193 0.001  
##      masse_ordma_min masse_ordma_max      champ_v9 date_maj Carrosserie
```

```
## 1      1505      1505 715/2007*692/2008EURO5 mars-14 BERLINE
## 2      1555      1555 715/2007*692/2008EURO5 mars-14 BERLINE
## 3      1565      1565 715/2007*692/2008EURO5 mars-14 BERLINE
## 4      1565      1565 715/2007*692/2008EURO5 mars-14 BERLINE
## 5      1565      1565 715/2007*692/2008EURO5 mars-14 BERLINE
## 6      1565      1565 715/2007*692/2008EURO5 mars-14 BERLINE
##      gamme  X X.1 X.2 X.3
## 1 MOY-SUPER NA  NA  NA  NA
## 2 MOY-SUPER NA  NA  NA  NA
## 3 MOY-SUPER NA  NA  NA  NA
## 4 MOY-SUPER NA  NA  NA  NA
## 5 MOY-SUPER NA  NA  NA  NA
## 6 MOY-SUPER NA  NA  NA  NA
```

```
length(t)
```

```
## [1] 30
```

Sur cette thématique, nous allons chercher à répondre aux deux questions suivantes :

- La masse d'une voiture est-elle corrélée à sa consommation ?
- Peut-on considérer que les voitures de gamme luxueuse rejettent autant de CO_2 que les voitures de gamme inférieure ?

I) Corrélation entre la masse et la consommation d'une voiture

Les données que nous utilisons nous donnent la possibilité de considérer plusieurs types de consommation :

- Urbaine
- Extra urbaine
- Mixte

Cependant, nous nous sommes intéressés à la consommation extra urbaine, qui nous paraissait plus pertinente.

Nous cherchons donc à comparer deux séries de données quantitatives, donc nous avons utilisé un test de corrélation, ayant pour hypothèse H_0 que les deux séries sont décorrélées.

```
cor.test(t$conso_exurb, t$masse_ordma_max)
```

```
##
## Pearson's product-moment correlation
##
## data:  t$conso_exurb and t$masse_ordma_max
## t = 151.79, df = 18781, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7357561 0.7486019
## sample estimates:
##      cor
## 0.7422472
```

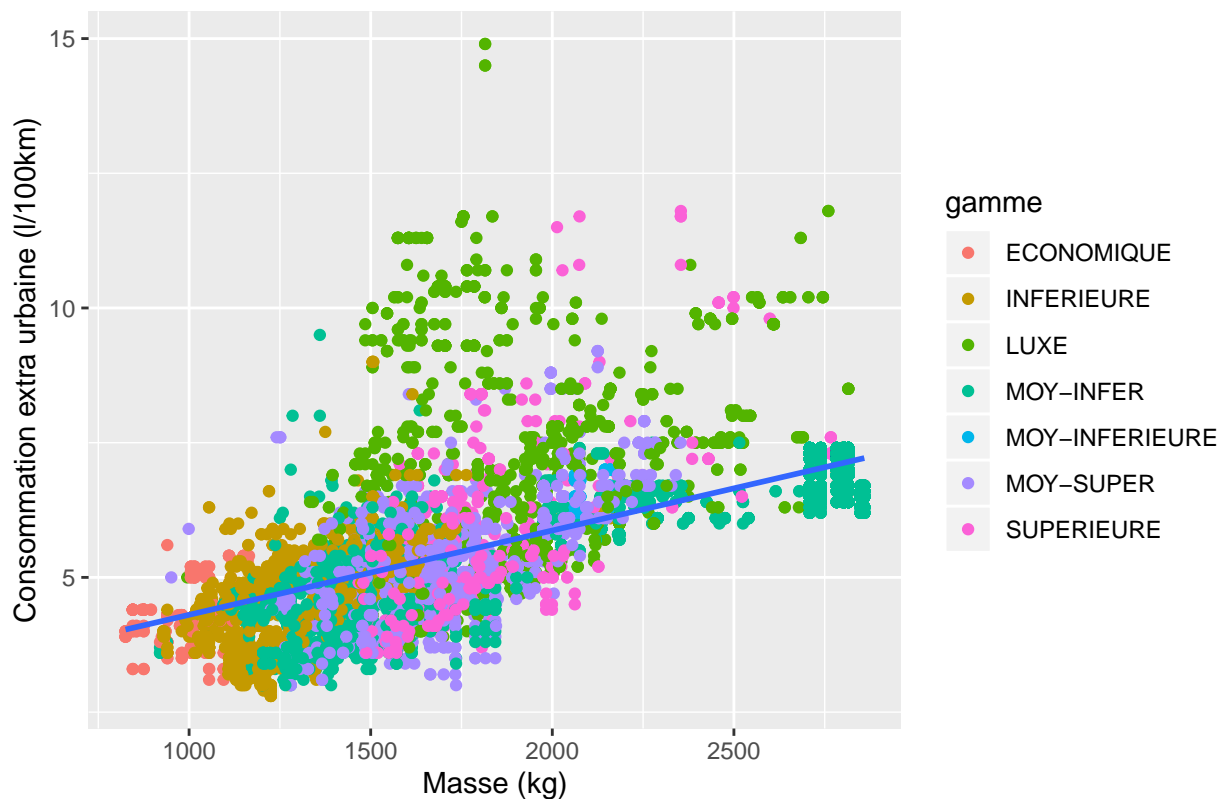
On trouve une p-valeur inférieure à 2.2×10^{-16} , donc au risque $\alpha = 5\%$, nous pouvons rejeter l'hypothèse H_0 .

Nous pouvons ainsi conclure (au risque $\alpha = 5\%$) que la masse de la voiture est corrélée à sa consommation extra urbaine.

De plus, on obtient un coefficient de corrélation de 0.74, ce qui nous induit à penser que le modèle linéaire n'est probablement pas le plus approprié.

```
gg <- ggplot(t, aes(x=masse_ordma_max, y=conso_exurb)) + geom_point(aes(col=gamme)) +  
  geom_smooth(method="lm") +  
  labs(title="Corrélation entre la masse d'une voiture et sa consommation extra urbaine") +  
  xlab("Masse (kg)") + ylab("Consommation extra urbaine (l/100km)")  
plot(gg)
```

Corrélation entre la masse d'une voiture et sa consommation extra urbaine



Sur le graphique ci-dessus, on remarque que les voitures de gamme économique et inférieure semblent consommer moins que celles de luxe, d'où la pertinence de la seconde question.

II) Comparaison du rejet de CO_2 pour différentes gammes

À la vue du graphique ci-dessus, nous avons choisi de comparer la gamme luxueuse à la gamme inférieure puisque c'est pour ces deux gammes que les différences semblent être les plus notables.

```

tinf <- t %>%
  filter(gamme == "INFERIEURE")

tluxe <- t %>%
  filter(gamme == "LUXE")

t2 <- t %>%
  filter(gamme == "LUXE" | gamme == "INFERIEURE")

head(t2)

##      lib_mrq lib_mod_doss  lib_mod                      dscom
## 1 ALFA-ROMEO          4C      4C                      4C
## 2 ALFA-ROMEO  AR8C SPIDER 8C SPIDER                      8C SPIDER
## 3 ALFA-ROMEO  AR8C SPIDER 8C SPIDER                      8C SPIDER
## 4 ALFA-ROMEO      MITO      MITO MITO 0.9 Twin Air (105ch) S/S
## 5 ALFA-ROMEO      MITO      MITO MITO 0.9 Twin Air (105ch) S/S
## 6 ALFA-ROMEO      MITO      MITO MITO 0.9 Twin Air (105ch) S/S
##      cnit      tvv cod_cbr hybride puiss_admin_98 puiss_max
## 1 M10ALFVP000S413  960CXB1A01      ES      non      14      177.0
## 2 M10ALFVP0005293  920BXA1A00B      ES      non      38      331.0
## 3 M10ALFVP0006039  920BXA1A00      ES      non      38      331.0
## 4 M10ALFVP000J379  955AXY1B18      ES      non      5       77.0
## 5 M10ALFVP000K380  955AXY1B18B      ES      non      5       77.0
## 6 M10ALFVP000U451  955AXZ1B21      ES      non      5       73.5
##      typ_boite_nb_rapp conso_urb conso_exurb conso_mixte co2 co_typ_1  hc  nox
## 1      A 6      9.8      5.0      6.8 157  0.404 0.044 0.038
## 2      M 6     24.4     11.6     16.3 379  0.501 0.038 0.027
## 3      M 6     24.4     11.6     16.3 379  0.501 0.038 0.027
## 4      M 6      5.0      3.8      4.2 99  0.316 0.041 0.034
## 5      M 6      5.0      3.8      4.2 99  0.316 0.041 0.034
## 6      M 6      5.0      3.8      4.2 88  0.316 0.041 0.034
##      hcnox  ptcl masse_ordma_min masse_ordma_max      champ_v9 date_maj
## 1      NA 0.003      995      995 715/2007*195/2013EUR06  mars-14
## 2      NA  NA      1750      1750 715/2007*692/2008EUR05  mars-14
## 3      NA  NA      1750      1750 715/2007*692/2008EUR05  mars-14
## 4      NA  NA      1205      1205 715/2007*630/2012EUR06  mars-14
## 5      NA  NA      1205      1205 715/2007*630/2012EUR06  mars-14
## 6      NA  NA      1205      1205 715/2007*195/2013EUR06  mars-14
##      Carrosserie      gamme X X.1 X.2 X.3
## 1      COUPE      LUXE NA  NA  NA  NA
## 2  CABRIOLET      LUXE NA  NA  NA  NA
## 3  CABRIOLET      LUXE NA  NA  NA  NA
## 4  BERLINE INFERIEURE NA  NA  NA  NA
## 5  BERLINE INFERIEURE NA  NA  NA  NA
## 6  BERLINE INFERIEURE NA  NA  NA  NA

length(t2)

## [1] 30

```

Nous cherchons donc à relier la gamme, une donnée qualitative, au rejet de CO_2 , une donnée quantitative.

Nous allons donc mener un test de comparaison de moyenne entre les données des deux gammes.

Pour bien choisir le test, nous devons déjà vérifier si les données suivent une loi normale. Pour cela, nous avons utilisé un test de Shapiro, ayant pour hypothèse H_0 que les données suivent une loi normale.

```
shapiro.test(tinf$co2)

##
##  Shapiro-Wilk normality test
##
## data:  tinf$co2
## W = 0.97684, p-value = 5.787e-12

shapiro.test(tluxe$co2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  tluxe$co2
## W = 0.94416, p-value < 2.2e-16
```

Dans les deux cas, la p-valeur est inférieure à un risque $\alpha = 5\%$. On peut donc rejeter l'hypothèse de normalité.

Nous allons ainsi mener un test non paramétrique de Wilcoxon, ayant pour hypothèse H_0 que les moyennes des deux séries peuvent être considérées comme identiques, c'est-à-dire que la différence peut s'expliquer simplement avec les fluctuations d'échantillonnage.

```
wilcox.test(tinf$co2, tluxe$co2)

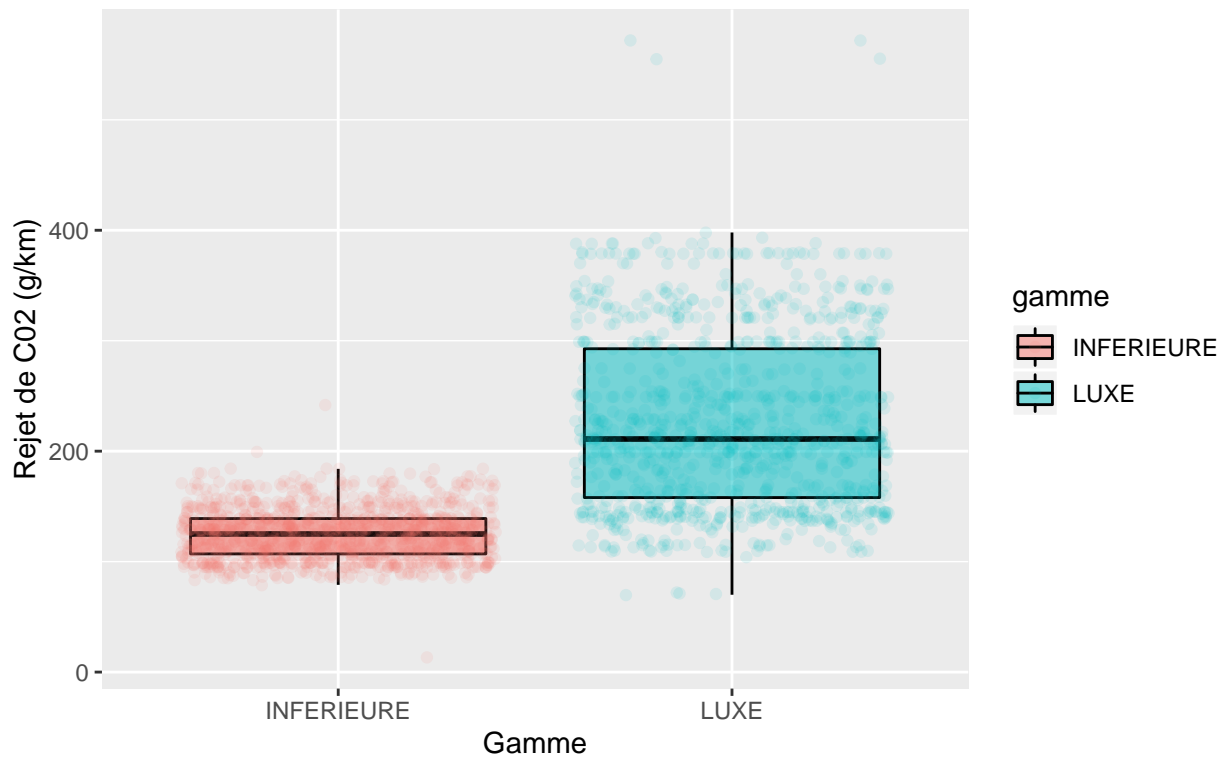
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  tinf$co2 and tluxe$co2
## W = 85238, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

On obtient ici aussi une p-valeur inférieure à 2.2×10^{-16} , donc pour un risque $\alpha = 5\%$ on peut rejeter H_0 .

Nous pouvons donc conclure (au risque $\alpha = 5\%$) que les voitures de luxe et les voitures de gammes inférieures ne consomment en moyenne pas autant.

```
gg <- ggplot(t2, aes(x=gamme, y=co2, fill=gamme, colour=gamme)) +
  geom_boxplot(alpha=0.5, outlier.alpha=0, colour="black") +
  geom_jitter(fill="black", alpha = 0.1) +
  labs(title="Rejet de CO2 pour les véhicules de gamme\n inférieure et de gamme luxueuse") +
  xlab("Gamme") + ylab("Rejet de CO2 (g/km)")
plot(gg)
```

Rejet de CO2 pour les véhicules de gamme inférieure et de gamme luxueuse



À la vue de nos résultats et du graphique ci-dessus, nous pouvons extrapoler que les voitures de luxe consomment en moyenne plus que les voitures de gamme inférieure, malgré une plus grande dispersion pour les voitures de luxe.

Conclusion

Nous avons pu répondre à la question de la corrélation entre masse et consommation à l'aide d'un test de corrélation, qui nous a révélé, au risque $\alpha = 5\%$, que ces deux paramètres sont effectivement corrélés.

À l'aide d'un test de Wilcoxon et en extrapolant le graphique associé, nous avons aussi conclu que les voitures de gammes inférieures consomment en moyenne moins que celles de gamme luxueuse.

Il ne faut cependant pas oublier que d'autres paramètres peuvent entrer en jeu, et que les liens ne sont pas nécessairement des liens de cause à effet.

Bibliographie

Les données utilisées (cliquez pour ouvrir).