Introduction

Chap. 1&2&3 du poly de cours - parties descriptives

Frédérique Leblanc

Utiliser les statistiques ? ... pour quoi ?

- Analyser des données disponibles sans information à priori ou question précise $\longrightarrow$ exploration de données (stats. desc.)
- Observer plusieurs répétitions d'une même expérience (dont l'issue est aléatoire) afin de répondre à une question précise (stats. inférentielles) avec des arguments statistiques (et significatifs)

Ex0 : les données suivantes sont disponibles dans R sous le nom `airquality` et donnent des mesures quotidiennes de qualité d'air à New York de May à Septembre 1973.

	Ozone	Solar.R	Wind	Temp	Month	Day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
5	NA	NA	14.3	56	5	5
.						
.						
.						

Ex1 : les données proposées dans les notes de cours collectées auprès d'une promotion d'étudiants de seconde année à l'UJF sont disponibles dans le fichier `donnees-polycfs-2016.csv`.

indiv	*sex*	*poids*20*a*	*poids*15*a*	*taille*	*alim*
1	1	49	45	160	5
2	1	53	45	164	4
3	1	50	45	161	4
4	1	49	45	175	6
...	...	...	...	...	...

Ex2 : données sur $n = 32$ véhicules : diverses caractéristiques des moteurs et du désign d'un véhicule, disponibles dans le data frame `mtcars` de R.
Ex3 : Diamètres et hauteurs de $n = 12$ arbres disponibles dans `arbres.csv`.

Diamètre	0.200	0.301	0.379	...	0.142
Hauteur	9.207	9.875	10.805	...	8.166

Ex4 : Autres données sur les performances techniques de véhicules

```
 > mpg
# A tibble: 234 x 11
   manufacturer      model displ  year   cyl      trans   drv   cty   hwy    fl   class
          <chr>      <chr> <dbl> <int> <int>      <chr> <chr> <int> <int> <chr>   <chr>
 1         audi         a4   1.8  1999     4   auto(l5)     f    18    29     p compact
 2         audi         a4   1.8  1999     4 manual(m5)     f    21    29     p compact
 3         audi         a4   2.0  2008     4 manual(m6)     f    20    31     p compact
 4         audi         a4   2.0  2008     4   auto(av)     f    21    30     p compact
 5         audi         a4   2.8  1999     6   auto(l5)     f    16    26     p compact
 6         audi         a4   2.8  1999     6 manual(m5)     f    18    26     p compact
 7         audi         a4   3.1  2008     6   auto(av)     f    18    27     p compact
 8         audi a4 quattro   1.8  1999     4 manual(m5)     4    18    26     p compact
 9         audi a4 quattro   1.8  1999     4   auto(l5)     4    16    25     p compact
10         audi a4 quatuor   2.0  2008     4 manual(m6)     4    20    28     p compact
# ... with 224 more rows
> names(mpg)
 [1] "manufacturer" "model"        "displ"        "year"         "cyl"          "trans"
 [7] "drv"          "cty"          "hwy"          "fl"           "class"
```

- Quelles sont les variables observées
- De quelle nature sont ces variables (qualitatives, quantitatives, discrètes, continues ...)?
- Les données sont-elles complètes ?

Dans tous les cas : quelles sont les variables, leurs unités,...
pour l'exemple 4 : l'aide de R renseigne sur les variables collectées et retourne

- cty : nombre de kilomètres en ville par gallon de carburant
- hwy : nb de km sur route par gallon
- drv : indique le type de traction (à valeurs dans $\{f, r, 4\}$)
- disp : cylindrée en litres...

De façon générale il s'agira de décrire les données et ensuite de valider statistiquement les hypothèses posées par le client ou suggérées par la description des données.
Par exemple dans le cas des données de EX0 (qualité de l'air) :

- Le taux d'Ozone suit-il une loi normale ?
- Le mois a t-il un effet sur le taux d'Ozone ?
- Observe-ton une différence significative entre le week-end et la semaine ?
- L'observatoire de la santé indique qu'au delà d'un seuil s_0 il est préférable de ne pas faire de sport en extérieur : peut-on considérer ce seuil dépassé courant Juin ?

Dans le cas de EX 4, on pourrait s'interroger sur

- l'effet du nombre de cylindres sur la consommation
- lien entre la consommation sur route ou en ville
- lien entre disp et hwy selon un critère comme cyl..

Ce qu'il faut faire :

- Décrire (nature des variables, ens. de leurs valeurs poss., résumés graphiques et numériques)
- Modéliser : proposer une loi pour la variable étudiée, caractérisée par un ou quelques paramètres inconnus (modèles paramétriques)
- Ajuster les paramètres du modèle à l'aide des données (estimation)
- Valider le modèle ajusté (vérification des hypothèses posées sur le modèle)
- Estimer, prendre un décision ou prévoir ...

1. 6 variables : 4 quantitatives continues et 2 qualitatives
   - Ozone : nombre moyen de particules d'Ozone sur un billion de particules (à un endroit et une heure précise)
   - Solar.R : mesure de radiation solaire
   - Wind : vitesse du vent
   - Temp : température (degr fahr.)
   - Month : mois
   - Day : jour
2. 5 mois d'observations quotidiennes
3. Données manquantes

Resumés Numériques :

```
> summary(data)
```

Ozone	Solar.R	Wind	Temp	Month
Min. : 1.00	Min. : 7.0	Min. : 1.700	Min. :56.00	Min. :5.000
1st Qu.: 18.00	1st Qu.:115.8	1st Qu.: 7.400	1st Qu.:72.00	1st Qu.:6.000
Median : 31.50	Median :205.0	Median : 9.700	Median :79.00	Median :7.000
Mean : 42.13	Mean :185.9	Mean : 9.958	Mean :77.88	Mean :6.993
3rd Qu.: 63.25	3rd Qu.:258.8	3rd Qu.:11.500	3rd Qu.:85.00	3rd Qu.:8.000
Max. :168.00	Max. :334.0	Max. :20.700	Max. :97.00	Max. :9.000
NA's : 37.00	NA's : 7.0			

Tableau de répartition d'une variable discrète :
Considérons la variable Month du data frame `aiquality` à valeurs dans l'ensemble de modalités $\{m_1, ..., m_5\}$:
Son tableau en effectifs :

m_k	5	6	7	8	9
n_k	31	30	31	31	30
f_k	0.2026144	0.1960784	0.2026144	0.2026144	0.1960784
F_k	0.2026144	0.3986928	0.6013072	0.8039216	1.0000000

n_k : effectifs, f_k : fréquences et F_k : fréquences cumulées

Resumés Graphiques pour une variable discrète :

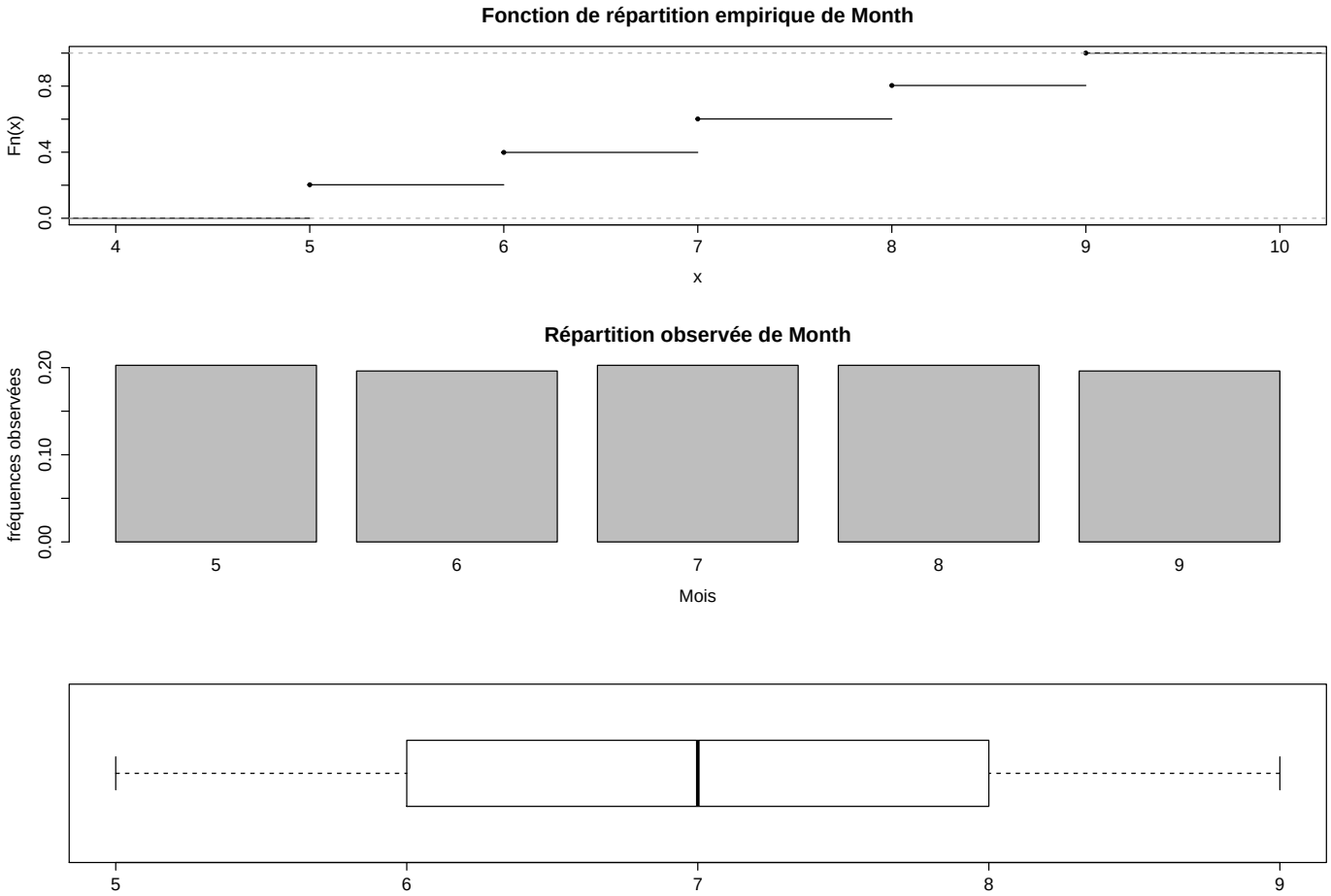


Tableau de répartition d'une variable continue :

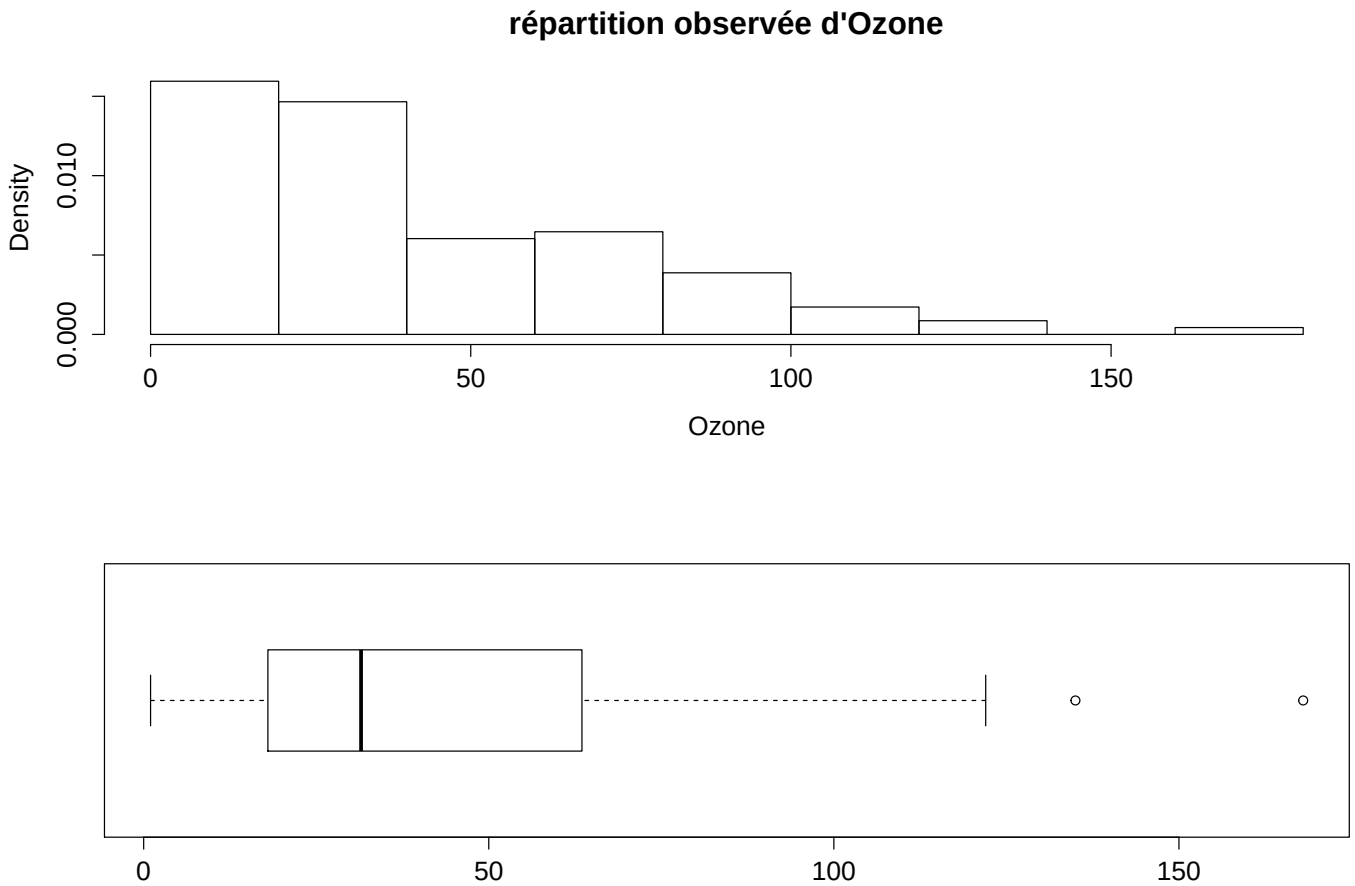
Considérons la variable Ozone du data frame `aiquality` à valeurs dans l'ensemble de classes $\{C_1, ..., C_q\}$ avec $C_k =]e_{k-1}, e_k]$:

Son tableau en effectifs :

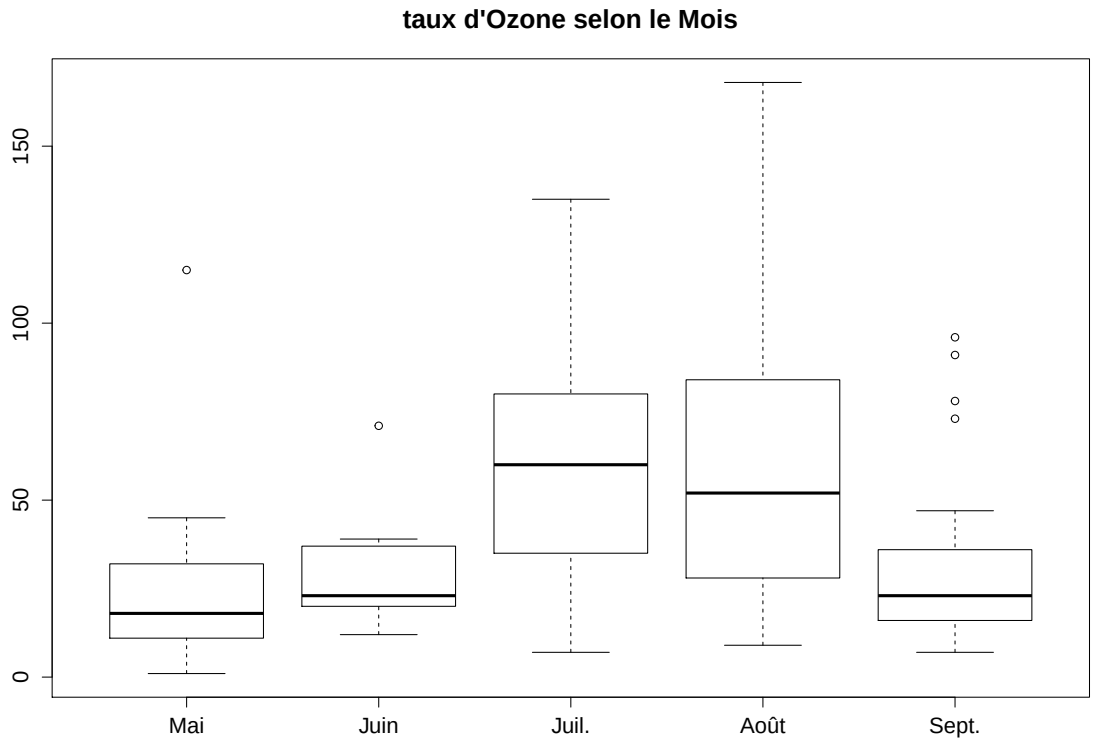
Classes	$]0; 20]$	$]20; 40]$	...	$]140; 160]$	$]160; 180]$
m_k	10	30	...	150	170
n_k	37	34	...	0	1
f_k	0.32	0.29	...	0.00	$9 \cdot 10^{-3}$
$f_k/(e_k - e_{k-1})$	0.016	0.015	...	0.00	$4 \cdot 10^{-4}$
F_k	0.32	0.61	...	0.99	1

m_k : milieux de classes, n_k:effectifs, f_k : fréquences et F_k : fréquences cumulées

Descriptions Graphiques d'une variable continue :

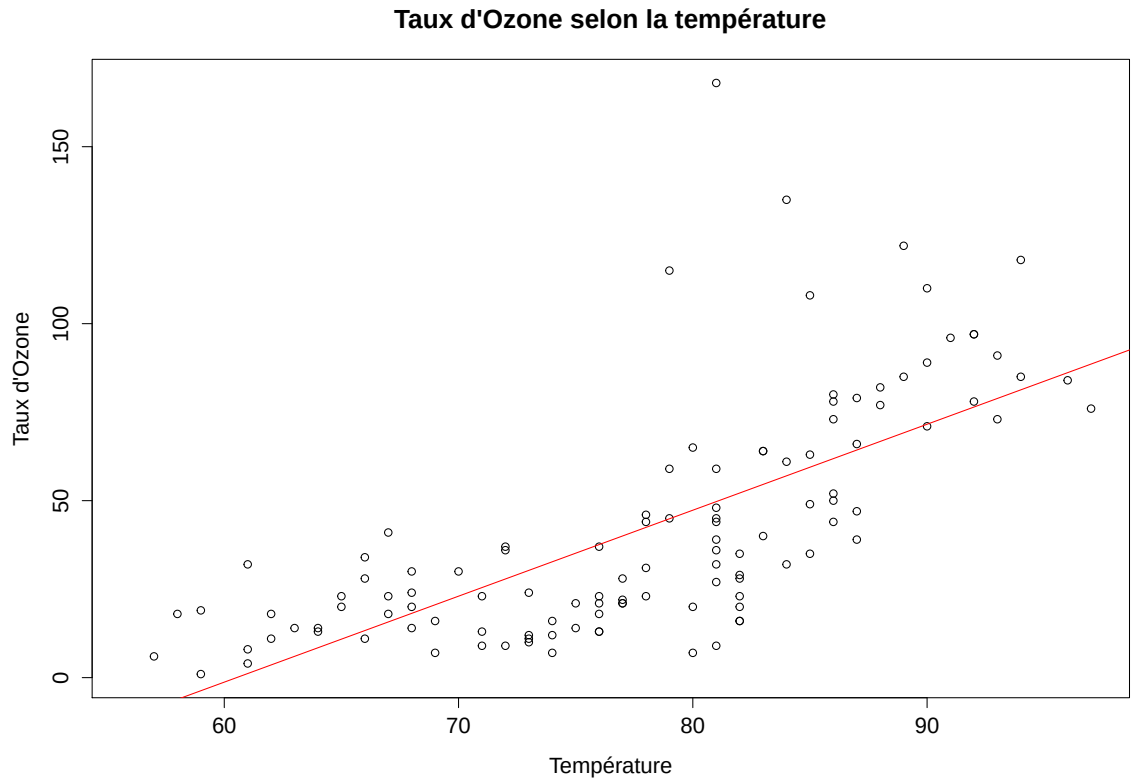


Quelques autres graphes descriptifs de `airquality`

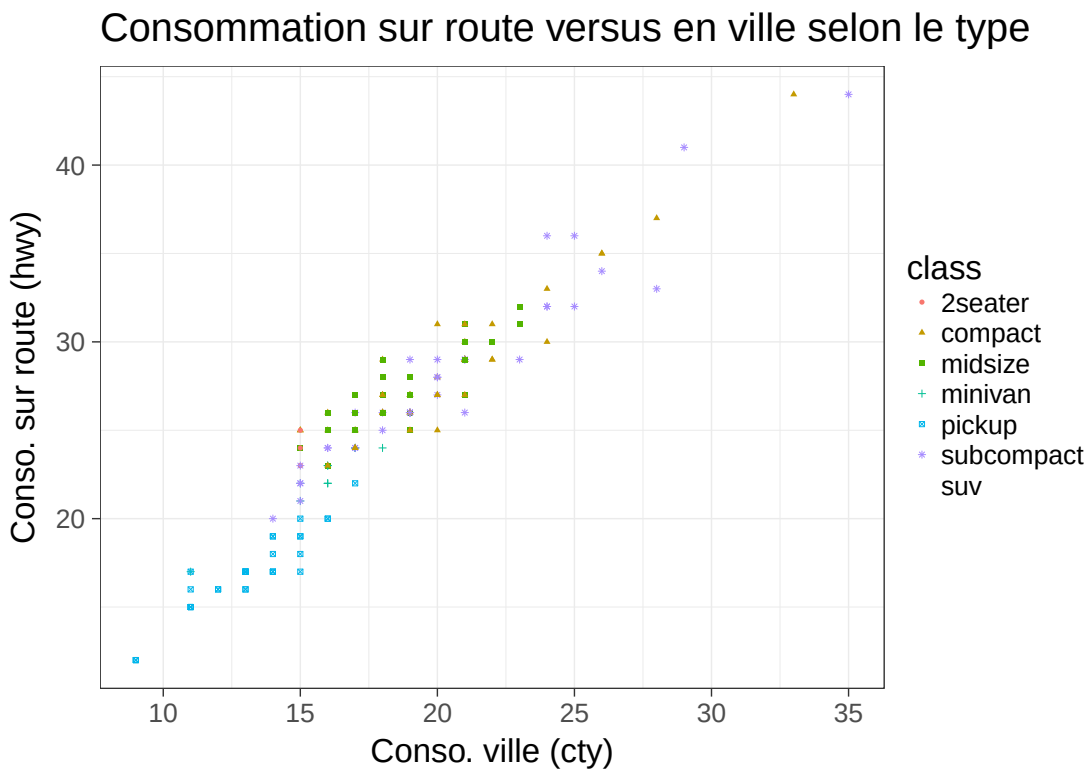


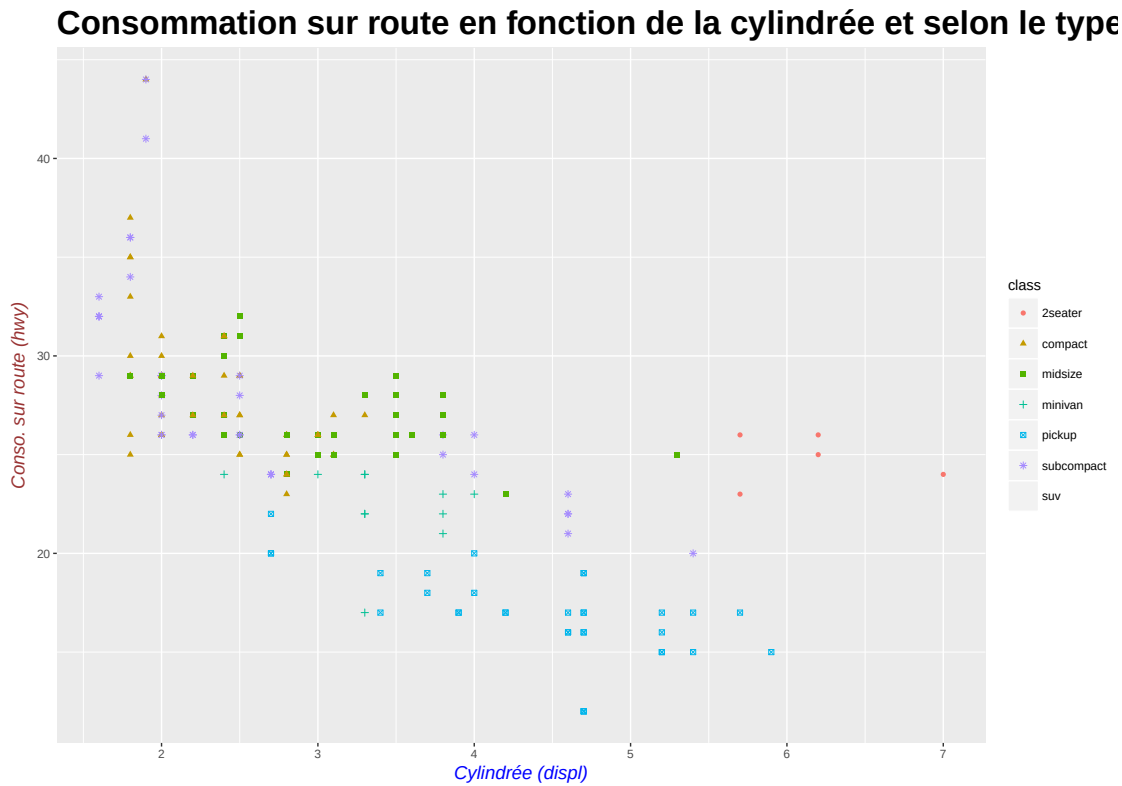
Suggère que le taux d'Ozone est plus élevé en Juillet et Août que les autres mois. Lien avec température ?

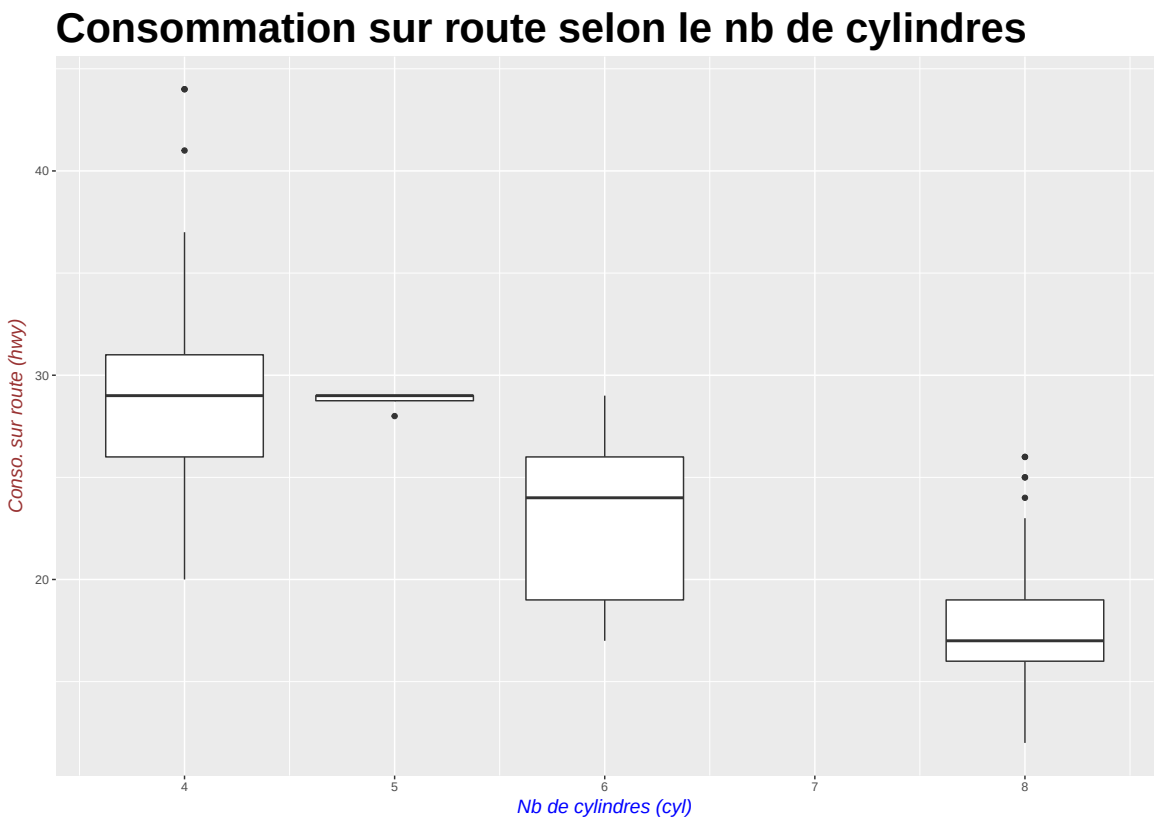
Et si on veut apprécier l'effet de la température sur le taux d'ozone on représente les points (t_i, o_i) pour les 153 mesures i



Pour les questions posées dans EX4 (mpg) des représentations graphiques adaptées peuvent donner des éléments de réponses :

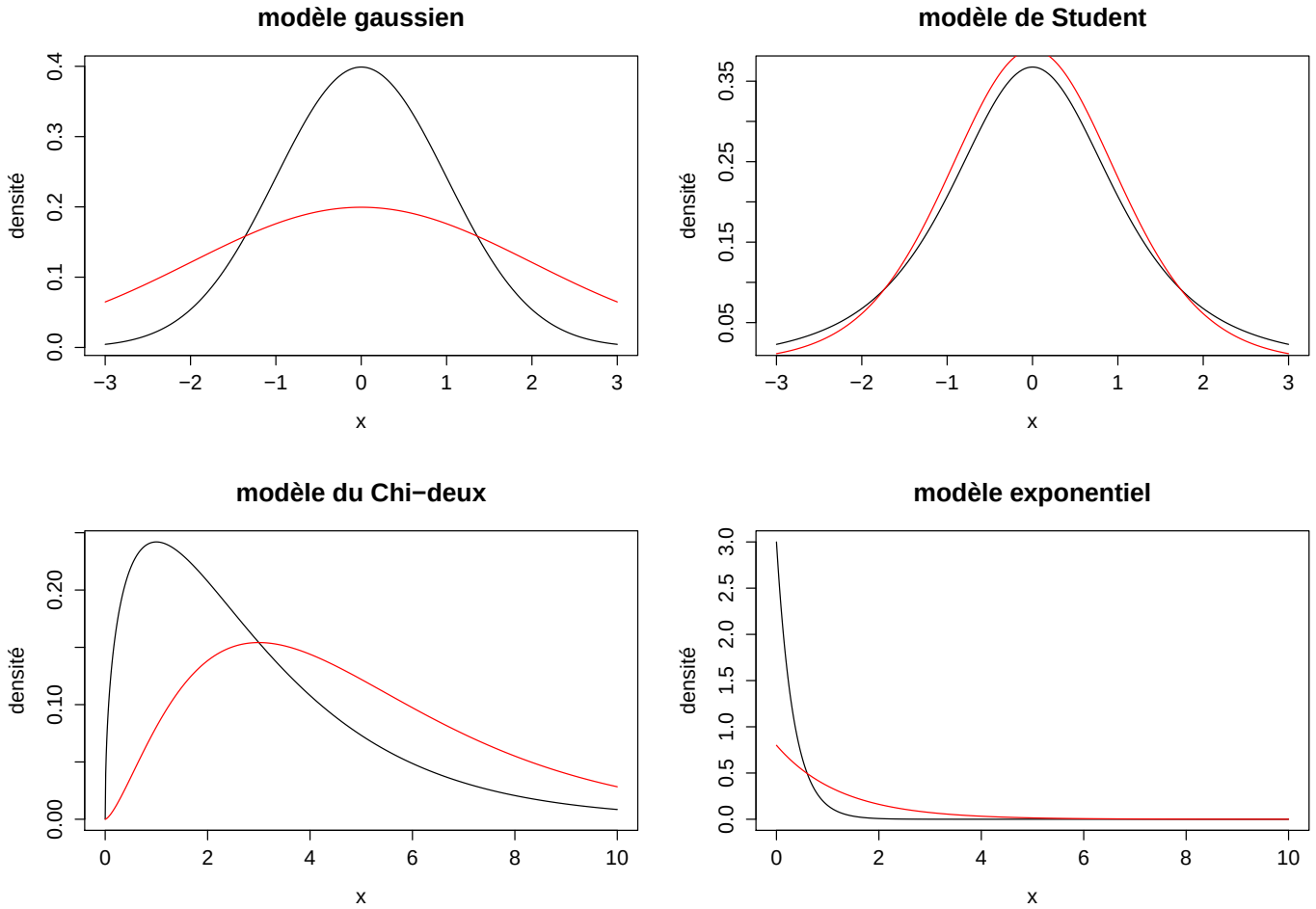


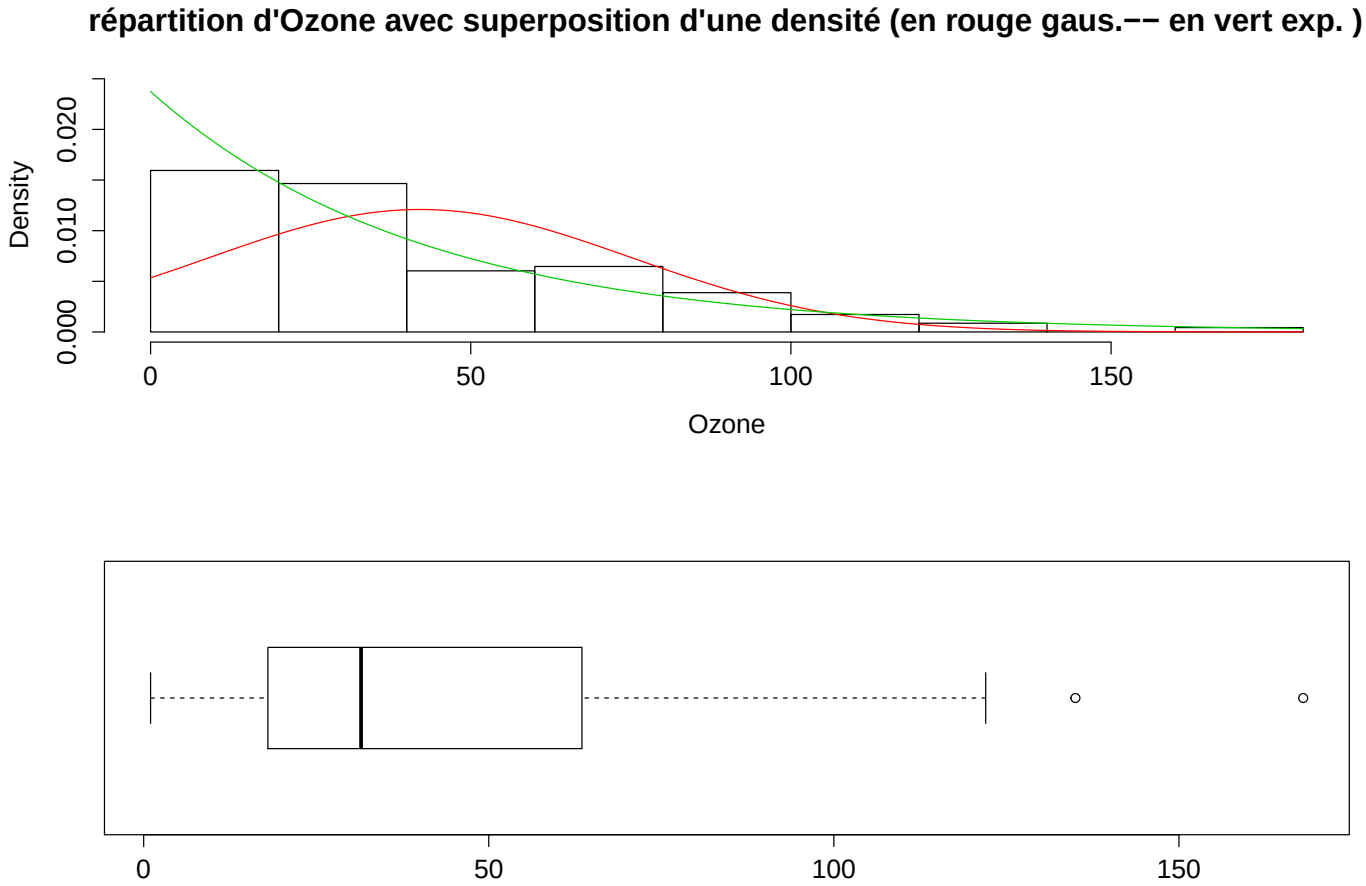

Consommation sur route en fonction de la cylindrée et selon le type
Conso. sur route (hwy)
40
30
20
2
3
4
5
6
7
Cylindrée (displ)
class
2seater
compact
midsize
minivan
pickup
subcompact
suv


Consommation sur route selon le nb de cylindres
Conso. sur route (hwy)
40
30
20
4
5
6
7
8
Nb de cylindres (cyl)

Trouver la famille de densités qui permettra d'ajuster au mieux la distribution observée parmi les lois continues disponibles :

- loi uniforme $\mathcal{U}([a, b])$
- loi normale $\mathcal{N}(\mu, \sigma^2)$
- loi du Chi-deux $\mathcal{X}_\nu$
- loi de Student $\mathcal{T}_\nu$
- loi exponentielle $\mathcal{E}(\lambda)$
- loi de Weibull.....


modèle gaussien
densité
x
modèle de Student
densité
x
modèle du Chi−deux
densité
x
modèle exponentiel
densité
x


répartition d'Ozone avec superposition d'une densité (en rouge gaus.−− en vert exp.)
Density
0.000
0.010
0.020
0
50
100
150
Ozone
0
50
100
150

- le modèle gaussien est moins adapté que le modèle exponentiel.
- tracer le graphe appelé `qqplot()` : quantiles empiriques (ord) vs quantiles théoriques du modèle cible (abs) et y ajouter la droite de Henry (d'équation $y = \sigma x + \mu$ dans le cas gaussien). Si les points sont alignés autour de la droite on jugera le modèle cible proposé adapté.
- avec R la fonction `qqnorm(x)` permet de vérifier graphiquement l'adéquation d'un modèle gaussien pour les données `x`. L'ajout de la droite de Henry se fera avec la commande `abline(mean(x),sd(x))`.

Dans le cas du taux d'Ozone on vérifie bien que le modèle exponentiel est meilleur. Néanmoins comme il s'agit d'un échantillon de grande taille on pourra faire "comme si" les données étaient gaussiennes grâce au TCL.