Exploration de données et Modélisation

Chap. 1&2&3 du polycopié

Chargés de cours
V. Léger &
F. Leblanc (resp. UE)

- Organisation :
  - 13 séances de cours, 12 séances de TD et 12 séances de TP
  - documents pédagogiques disponibles sur Chamilo : https://chamilo.univ-grenoble-alpes.fr/
- Evaluation : tout en contrôle continu avec 3 notes
  - $CC1$: note de TP (2 évaluations individuelles en séance)
  - $CC2$: note de TD (1 partiel de 1,5 h et 2 quizz de 15 min. en TD)
  - $CC3$: un examen terminal de 2 h

  $\text{Note Finale} = 20\%CC1+30\%CC2+50\%CC3 \quad \text{si} \quad CC2 \geq CC3$

  $\text{Note Finale} = 20\%CC1+0\%CC2+80\%CC3 \quad \text{si} \quad CC2 < CC3$

Rem : conservation possible de $CC1$ si $CC1 \geq 10$ pour les redoublants

Motivations et Objectifs

Les statistiques ... pour faire quoi ?

1. Analyser des données disponibles :
   - explorer et visualiser les données (stats. desc.)
   - interpréter et poser des conjectures
   - les valider (ou pas) à l'aide des outils de statistiques inférentielle
2. Répondre à une question précise (aide à la décision)
   - Collecter des données (observations répétées d'une même expérience)
   - Exploration et description des données
   - Mise en oeuvre d'un test pour répondre
   - Conclure avec des arguments statistiques significatifs (par ex. risque d'erreur dans les test)

Ex0 : `airquality`

Données disponibles dans R qui 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 : `donnees-poly.csv`

Données proposées dans les notes de cours collectées auprès d'une promotion d'étudiants de seconde année à l'UGA.

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 : `mtcars`
Diverses caractéristiques des moteurs et du désign d'un véhicule, disponibles dans R (étudiées dans le TP2)
Ex3 : Durées de vie d'une ampoule
Durée de vie de $n = 10$ ampoules (en heures) :

$$(x_1, ..., x_{10}) = (4.4, 2.6, 5.4, 7.8, 0.9, 0.5, 2.7, 9.1, 2.9, 1.2)$$

données ordonnées :

$$(x_{(1)}, ..., x_{(n)}) = (0.5, 0.9, 1.2, 2.6, 2.7, 2.9, 4.4, 5.4, 7.8, 9.1)$$

avec $x_{(1)}$ le minimum et $x_{(n)}$ le maximum

Ex4: mpg
Données sur les performances techniques de véhicules disponibles dans le package ggplot2 de R

```
 > 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"
```

EX5 : Notes d'examen avec données en effectifs

Soit X la note de $CC2$ de $n = 30$ étudiants $(x_1, ..., x_n)$ avec $X \in \mathcal{X} = \{1, ..., 20\}$ où $x_i \in \{m_1, ..., m_k\} = \{3, 5, 8, 9, 10, 11, 12, 14, 15, 16\}$ ensemble des modalités observées :

Grade (m_j)	3	5	8	9	10	11	12	14	15	16
Nb of students (n_j)	2	1	1	5	4	8	2	4	2	1

- Données sous forme d'un tableau en effectifs
- Cas discret : équivalent aux données brutes $(x_1, ..., x_n)$
- Cas continu $(\mathcal{X} = [0, 20])$: m_j centre de classe $=>$ perte d'information par rapport aux données brutes.

EX6 : Iris

Jeu de données R disponibles sous R dans le `data.frame` `iris` : 4 variables quantitatives continues (longueur et largeur de pétales et de sépale en cm) et une variable qualitative (espèce). Echantillon de $n = 50$ individus observés sur 5 variables.

Pour tous les jeux de données :

- Quelles sont les variables observées
- De quelle nature sont ces variables (qualitatives, quantitatives, discrètes, continues ...)?
- Les données sont-elles brutes ou en effectifs ?
- Variables manquantes (NA) ?

par ex pour EX2 : l'aide de R renseigne sur les variables collectées et retourne

- `cyl` : nombre de cylindres (cyl $\in \{4, 6, 8\}$) => cas discret car de nombreuses répétition dans $(x_1, ..., x_n)$
- `mpg` : nb de km par gallon d'essence (mpg $\in [10, 34]$) => cas cont. car peu de répétitions dans $(x_1, ..., x_n)$
- ...
- `am` : automatique ou manuelle (am $\in \{0, 1\}$ où 0 code la transmission automatique)

De façon générale il s'agira de décrire les données et ensuite de valider statistiquement les hypothèses posées par le client ou suggérées par l'exploration 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 EX4 (`mpg`), 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..

Démarche statistique :

- 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 et IC)
- Valider le modèle ajusté (vérification des hypothèses posées sur le modèle)
- Estimer, prendre une décision (test) ou prévoir ...

Revenons sur le jeu de données `airquality` de EX0

1. 6 variables : 4 quantitatives continues et 2 qualitatives
   - Ozone : nb moyen de particules d'Ozone sur un billion de particules => quanti. cont. (à un endroit et une heure précise) => var. cont.
   - Solar.R : mesure de radiation solaire => var. cont.
   - Wind : vitesse du vent => var. cont.
   - Temp : température (degr fahr.) => var. cont.
   - Month : mois => var. qualitative (facteur) **faussement traité ici comme quant. discrète pour illustrer les outils graphiques et résumés numériques**
   - Day : jour => var. qualitative (facteur)
2. 5 mois d'observations quotidiennes
3. Données manquantes

Resumés Numériques fournis par R du data frame `airquality` :

```
> 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	0.2026	0.1960	0.2026	0.2026	0.1960
F_k	0.2026	0.3986	0.6013	0.8039	1.0000

n_k les effectifs, $f_k = n_k/n$ les fréquences et $F_k = f_1 + ... + f_k$ les fréquences cumulées.

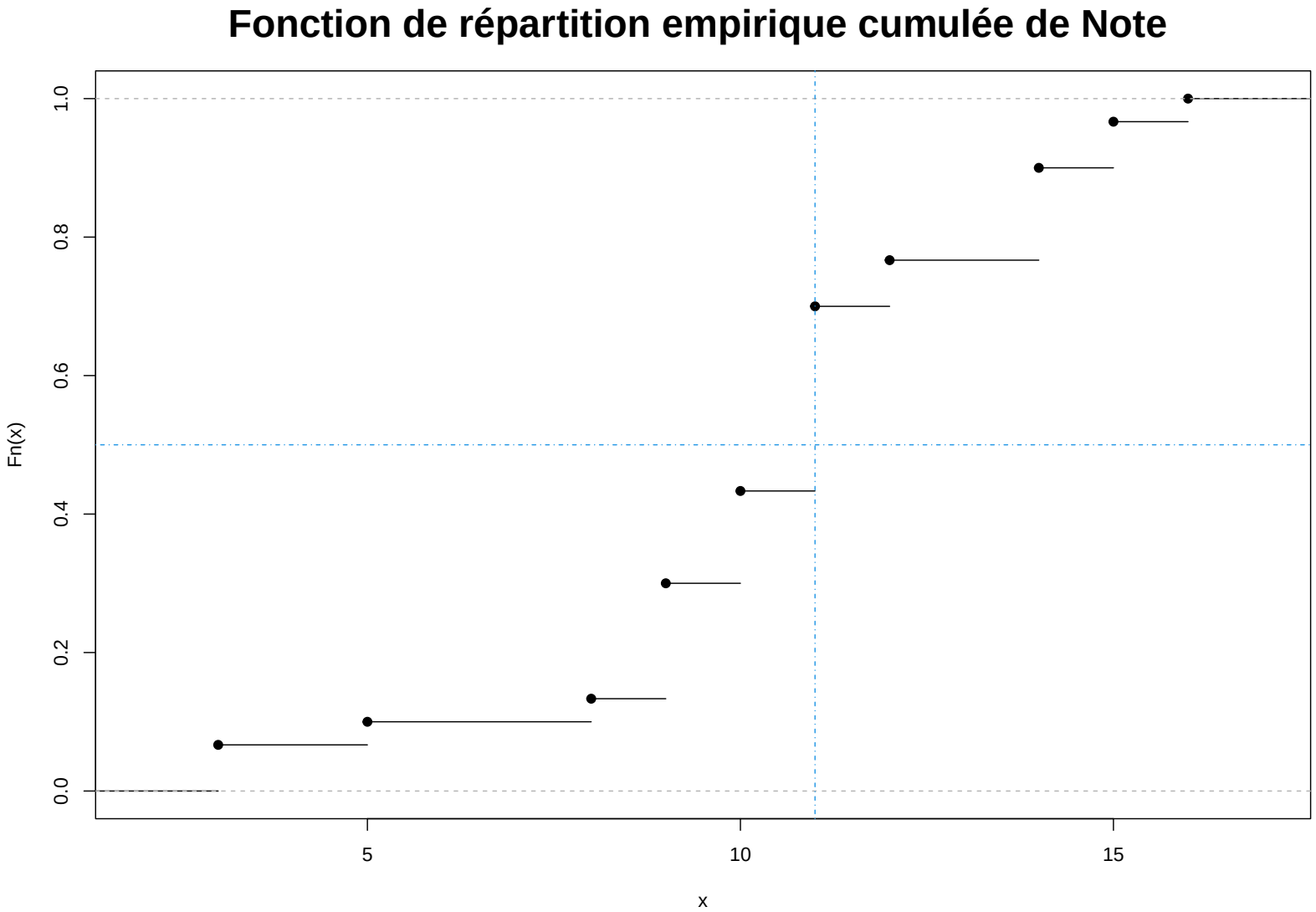
La fonction de répartition empirique cumulée

Pour un échantillon de données quantitatives (discrètes ou continues) c'est la fonction définie sur IR par :

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{x_i \leq x\}} = \frac{1}{n}\sum_{i=1}^{n} I_{\{x_{(i)} \leq x\}} = freq]-\infty, x]$$

- sous R : `ecdf` (empiricial cumulative distribution function)
- fonction en escaliers, croissante, valant 0 avant $x_{(1)}$ et 1 à partir de $x_{(n)}$
- dans le cas continu et avec des données en effectifs on peut l'approcher par une fonction continue, linéaire par morceaux (par ex dans EX1, voir Figure 3.3 du poly : var. poids à 20 ans)

La fonction de répartition cumulée pour les données de EX5



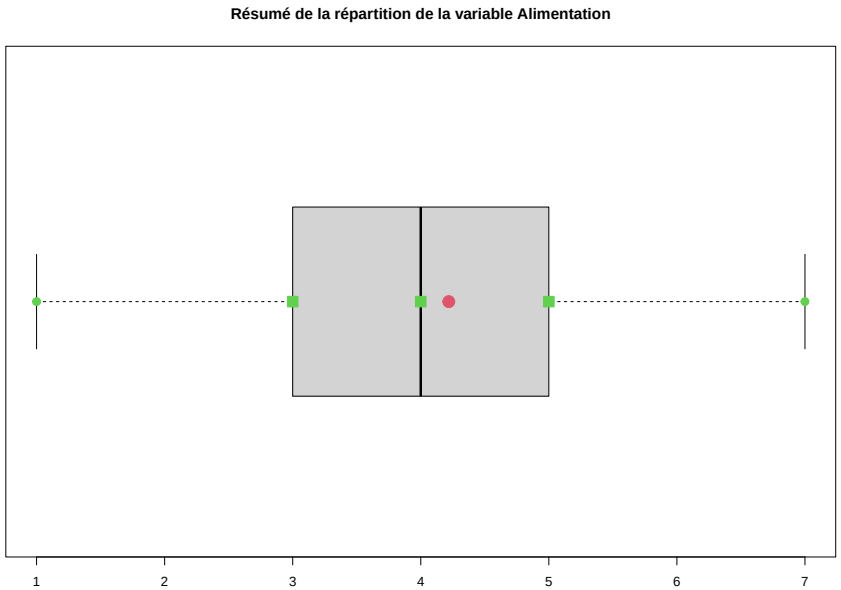
les résumés numériques (cas discret et continu modulo la définition de m_k):

- le quantile empirique d'ordre α : la première observation de $(x_{(1)}, ..., x_{(n)})$ pour laquelle $\hat{F}$ dépasse α c'est à dire la valeur $\hat{q}_\alpha$ telle que : $\hat{F}(\hat{q}_\alpha^-) < \alpha$ et $\hat{F}(\hat{q}_\alpha) \geq \alpha$
- les quartiles q_1, q_2 et q_3 sont les quantiles empiriques d'ordre respectifs 25%, 50% (médiane) et 75%
- la moyenne empirique :
  $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{k=1}^{q} n_k m_k = \sum_{k=1}^{q} f_k m_k$
- la variance empirique
  $s_n^2 = Q(\bar{x}_n) = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x}_n)^2 = \frac{1}{n}\sum_{k=1}^{q} n_k(m_k - \bar{x}_n)^2$
- l'écart-type empirique : s_n

Un premier résumé graphique : le box-plot ou boite à moustaches

- un résumé efficace de la répartition des données (discrètes ou continues)
- construit avec $q_1, q_2, q_3, min(x_i)$ et $max(x_i)$

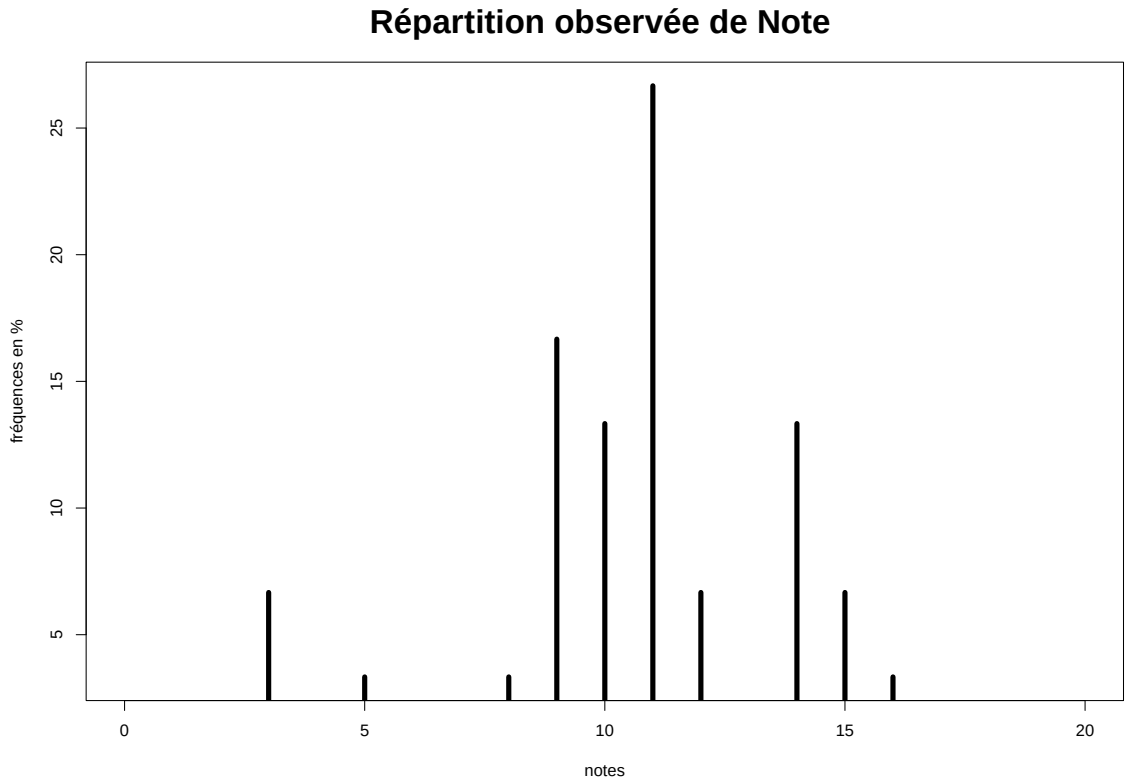
Par exemple celui de la variable `alim` de EX1



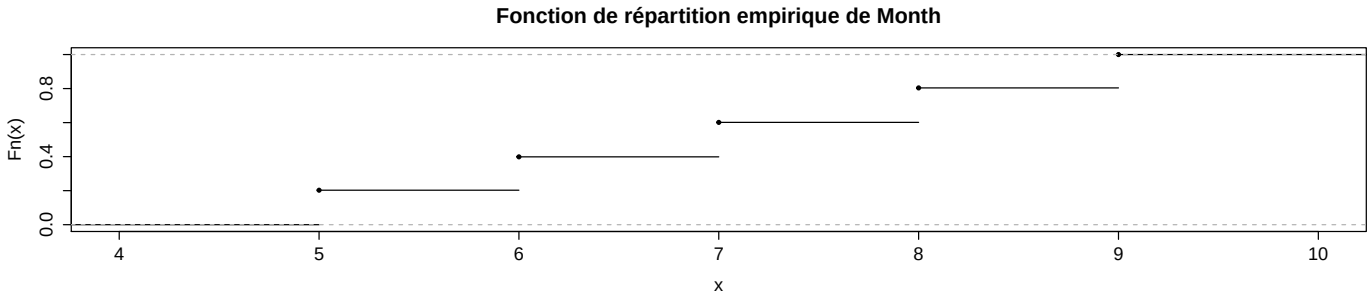
Les représentations graphiques pour des données discrètes

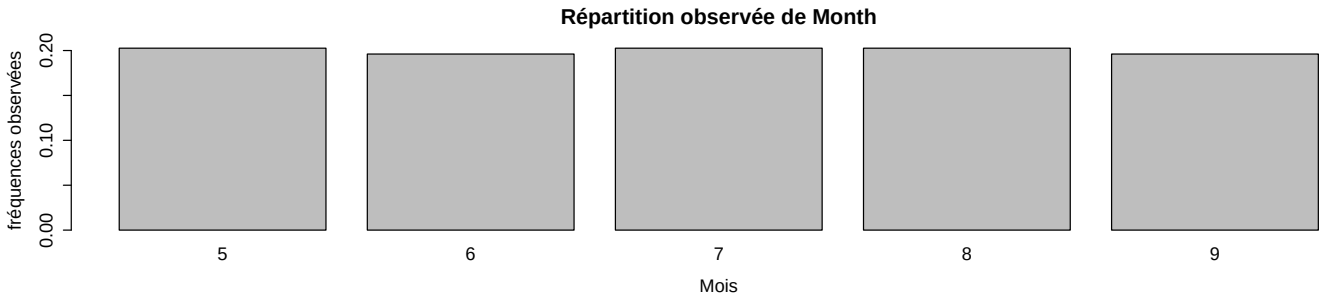
- Le box-plot (boite à moustaches ou boite de distribution)
- Le barplot : diagramme en barres où la hauteur de la barre pour la modalité m_k est $f_k = n_k/n$ et la barre placée en l'abscisse m_k
- La fonction de répartition empirique cumulée

Par exemple pour la variable Note de EX5 on a le barplot suivant (attention avec R, utiliser la fonction `plot` pour avoir un axe des abscisses avec une unité car la fonction `barplot` n'a pas d'unité en abscisse)



Pour la variable Month de EX0





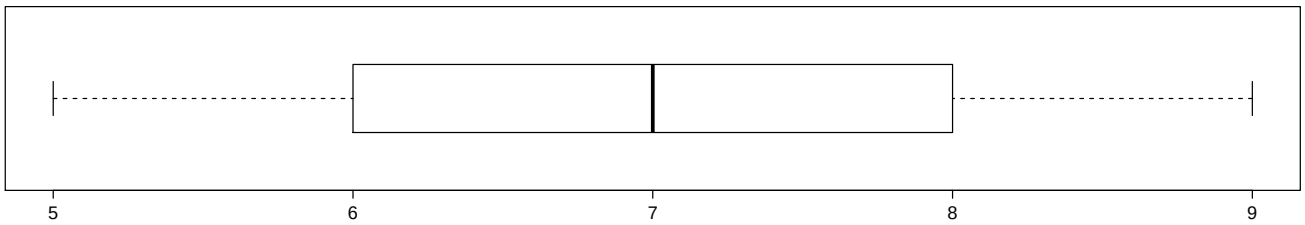


Tableau de répartition d'une variable continue :

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 :

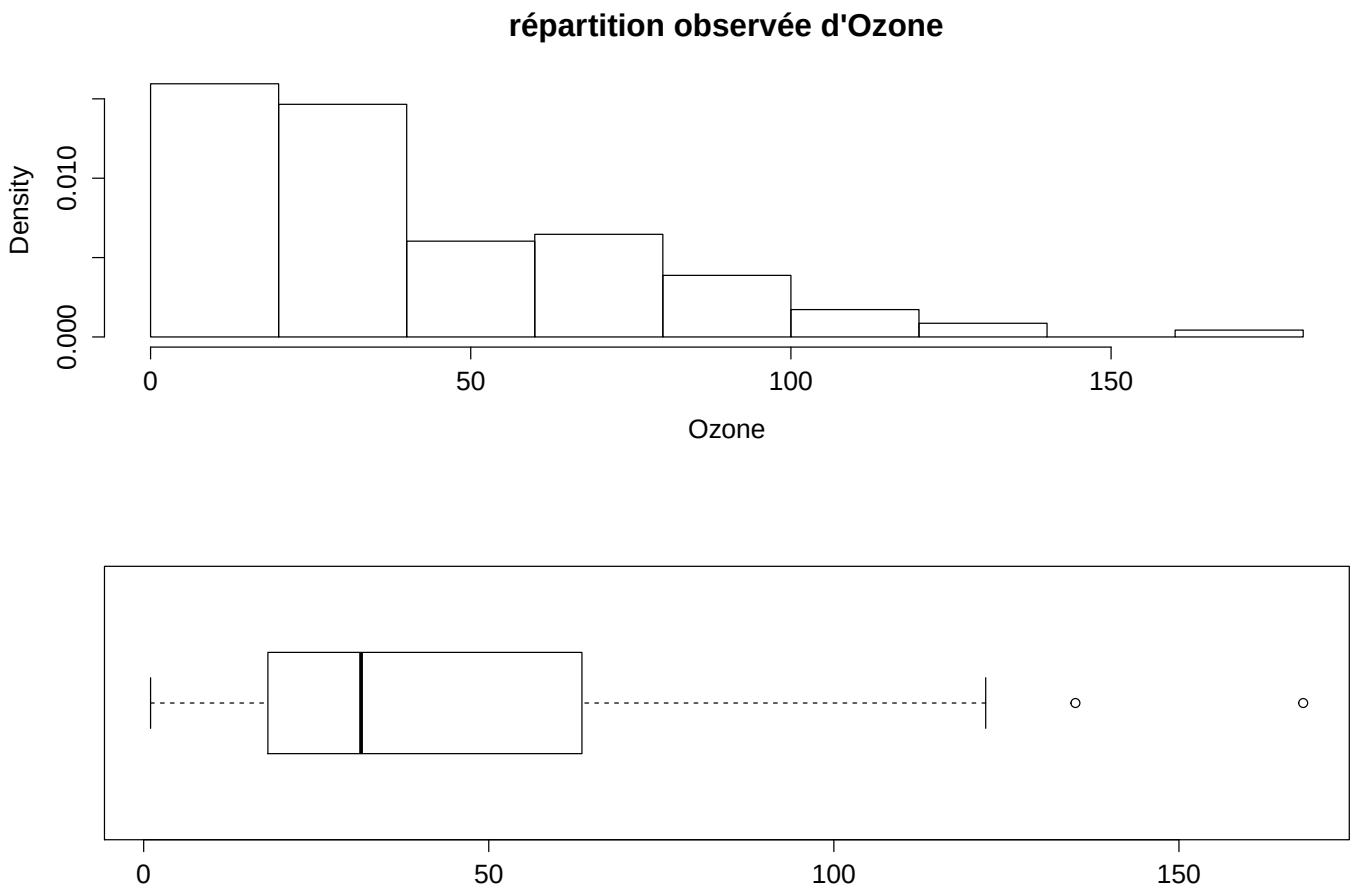
Classes	]0; 20]	]20; 40]	...	]140; 160]	]160; 180]
m_k	10	30	...	150	170
n_k	37	34	...	0	1
f_k	0.32	0.29	...	0.00	$9 \cdot 10^{-3}$
$f_k/(e_k - e_{k-1})$	0.016	0.015	...	0.00	$4 \cdot 10^{-4}$
F_k	0.32	0.61	...	0.99	1

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

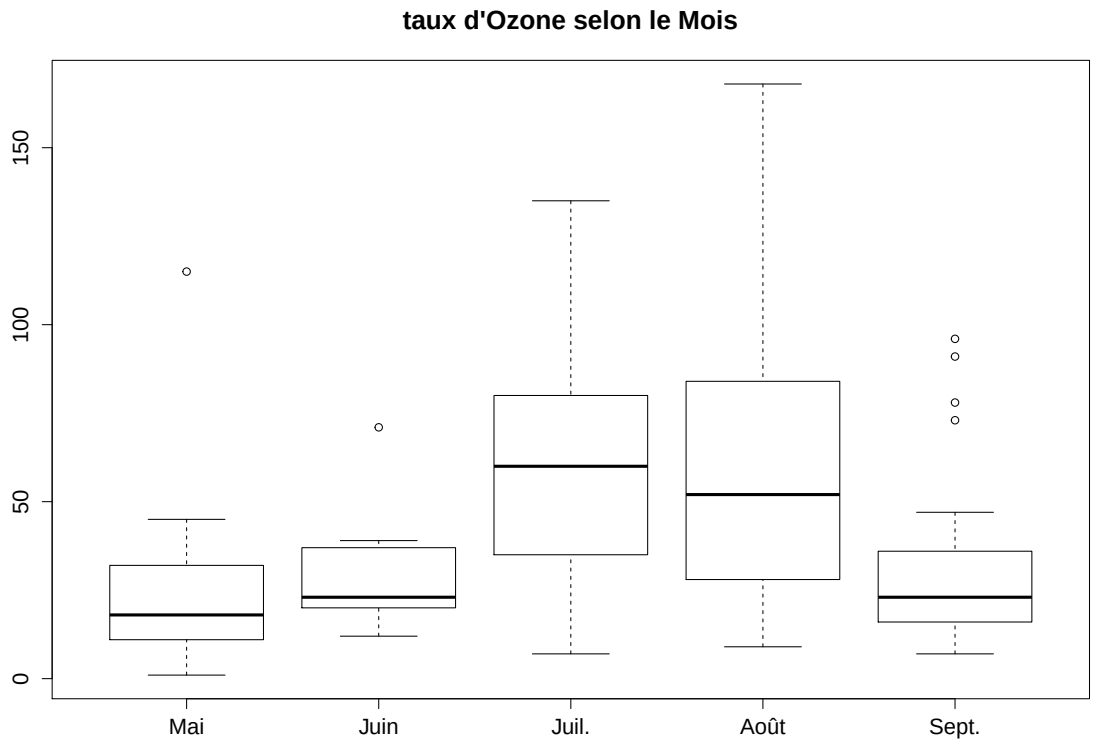
Les représentations graphiques pour des données continues

- Le box-plot (boite à moustaches ou boite de distribution)
- L'histogramme (en densité) : diagramme en "rectangles" où la hauteur du rectangle pour la classe k est $f_k/(e_k - e_{k-1})$ et sa largeur $(e_k - e_{k-1})$ (de surface f_k)
- La fonction de répartition empirique cumulée

Descriptions Graphiques de la variable Ozone de EX0 :

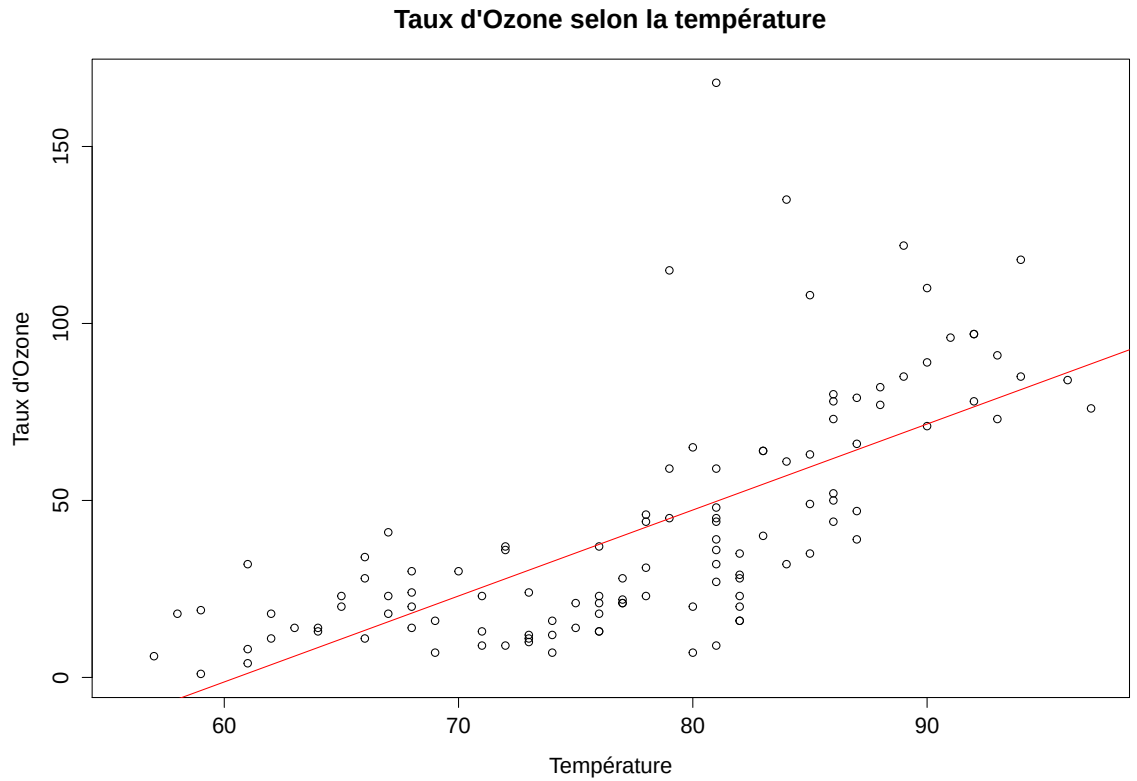


Quelques autres graphes descriptifs de `airquality`

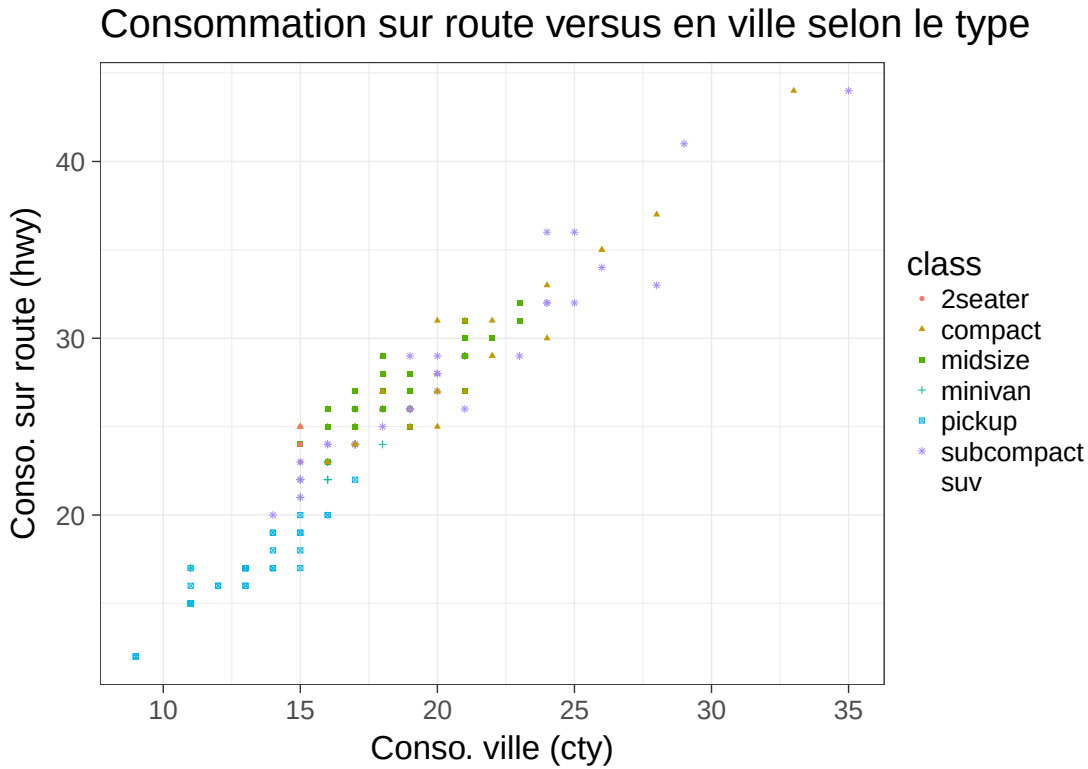


Suggère que le taux d'Ozone est plus élevé en Juillet et Août que les autres mois. Lien avec température ?

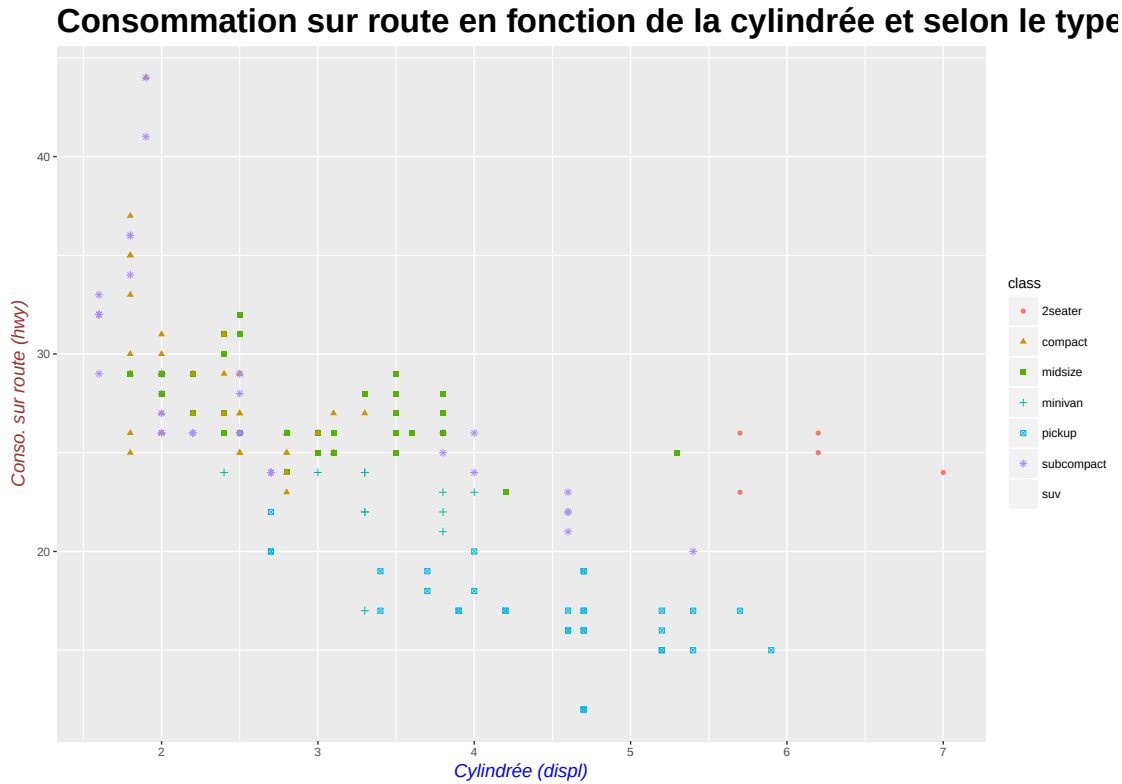
Et si on veut apprécier l'effet de la température sur le taux d'ozone on représente les points (t_i, o_i) pour les 153 mesures i

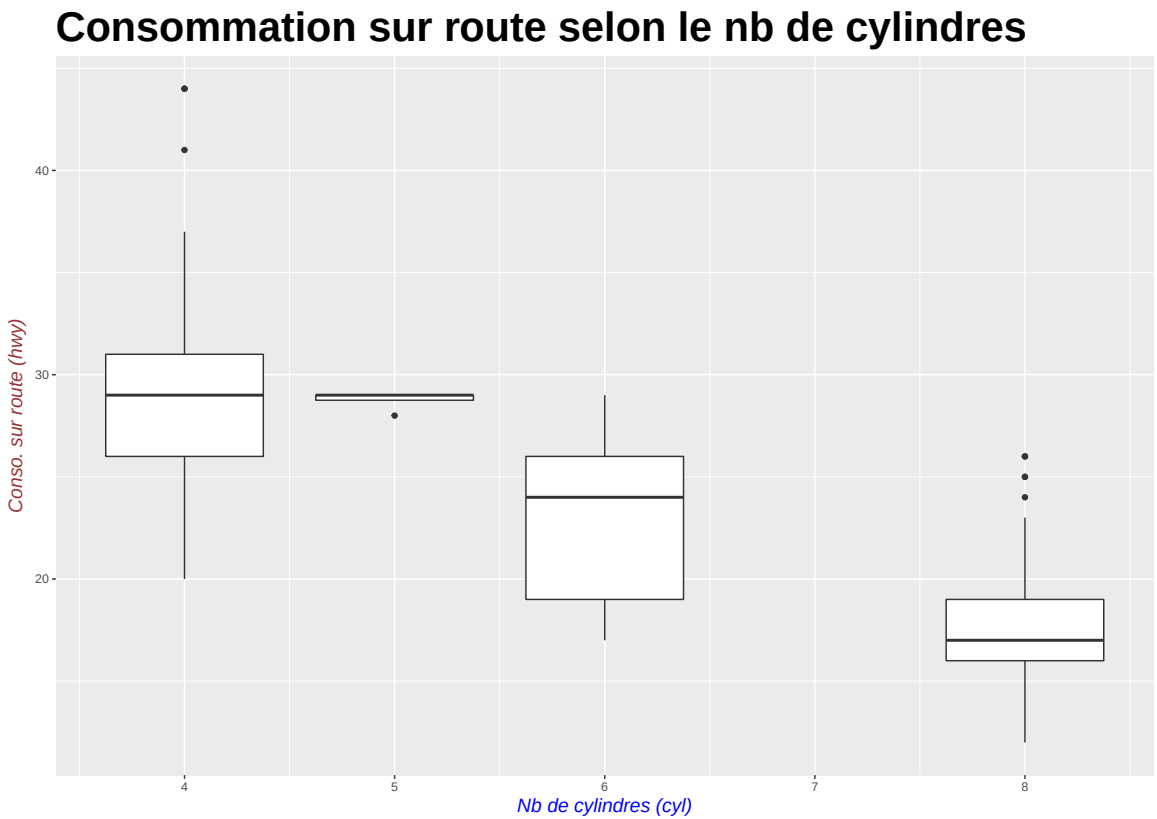


Pour les questions posées dans EX4 (mpg) des représentations graphiques adaptées peuvent donner des éléments de réponses. Par exemple pour visualiser l'effet de la classe de véhicule sur la corrélation entre `cty` et `hwy`

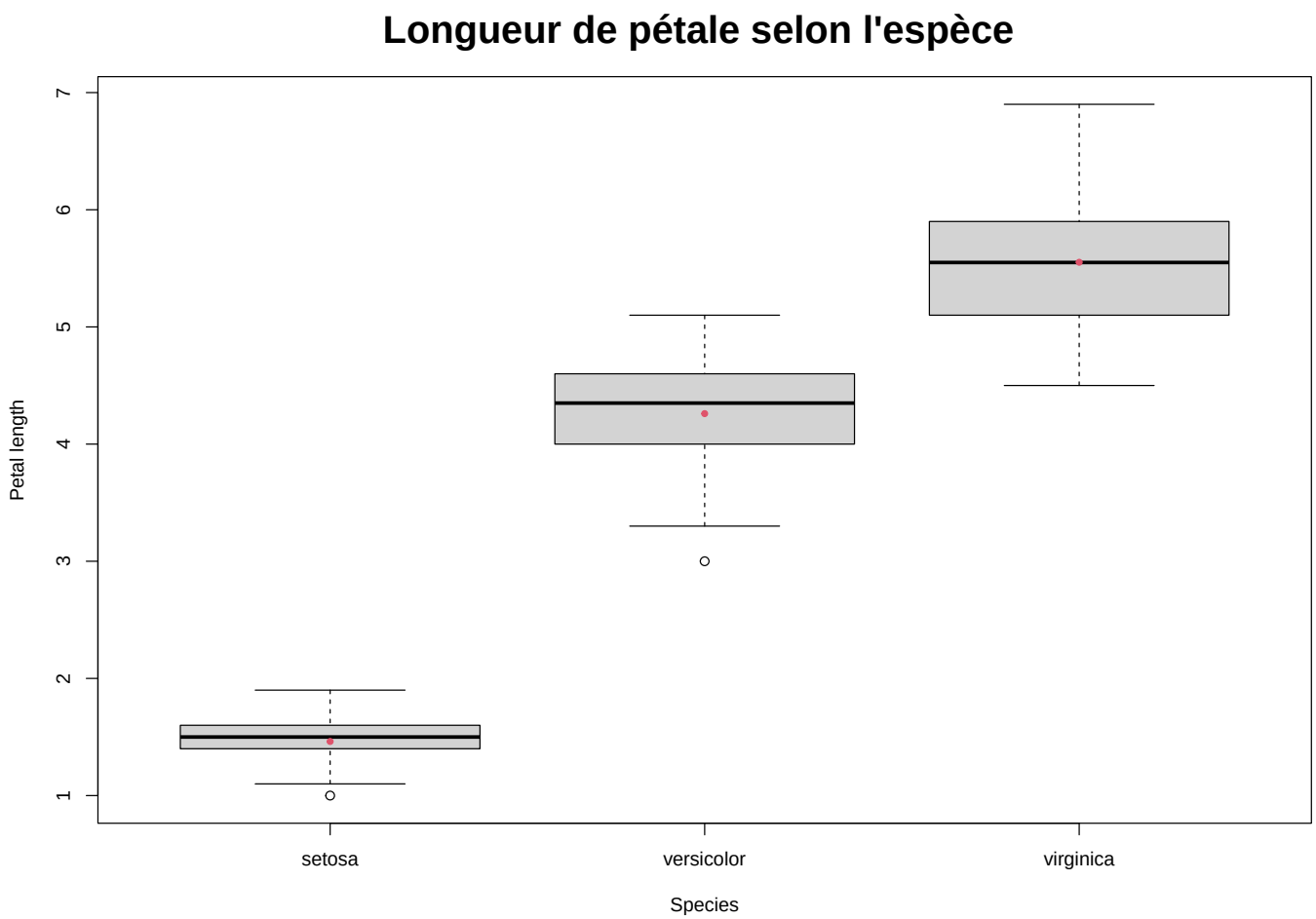


ou encore pour apprécier le lien entre les variables `disp` et `hwy`

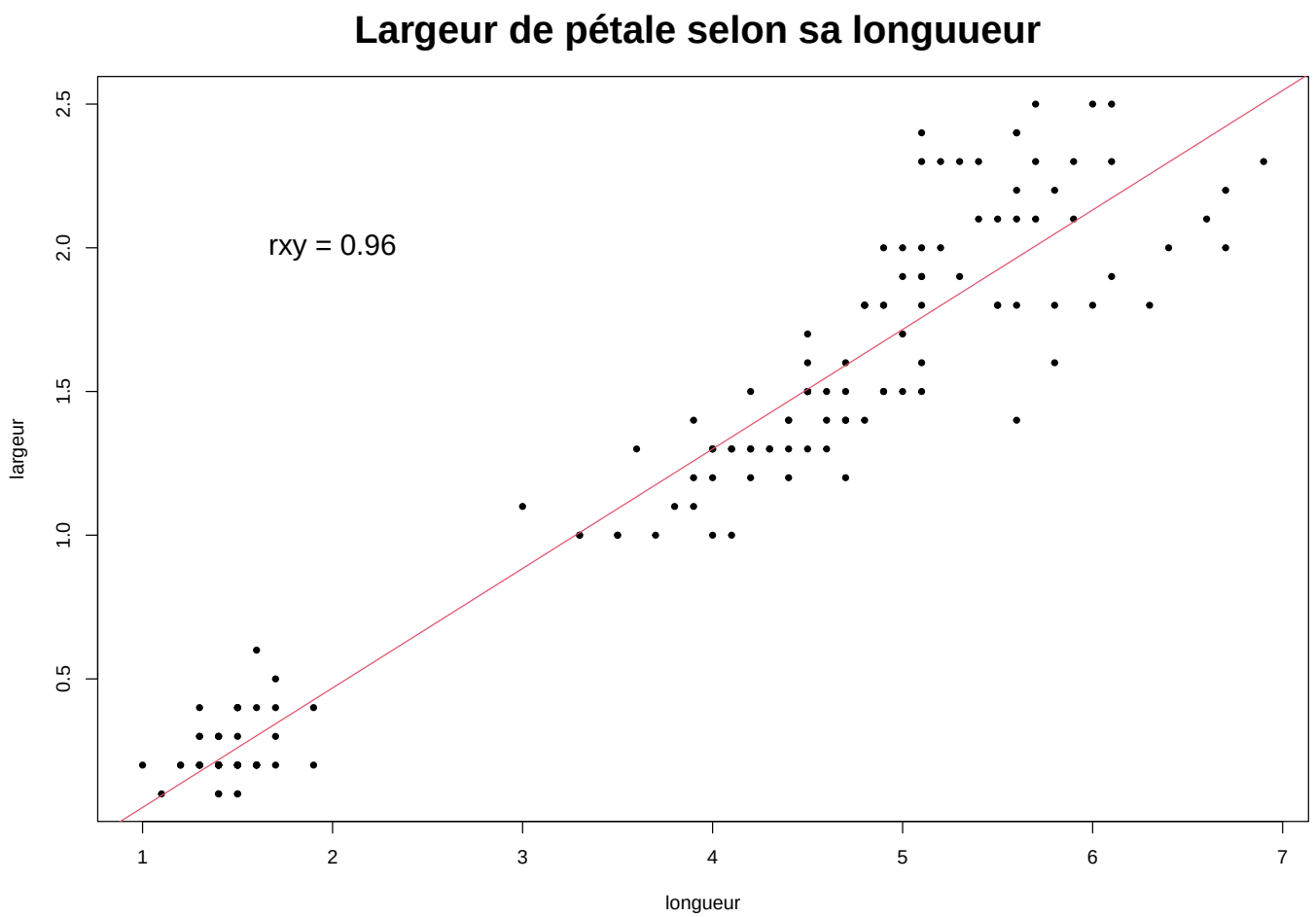



Consommation sur route selon le nb de cylindres
Conso. sur route (hwy)
40
30
20
4
5
6
7
8
Nb de cylindres (cyl)

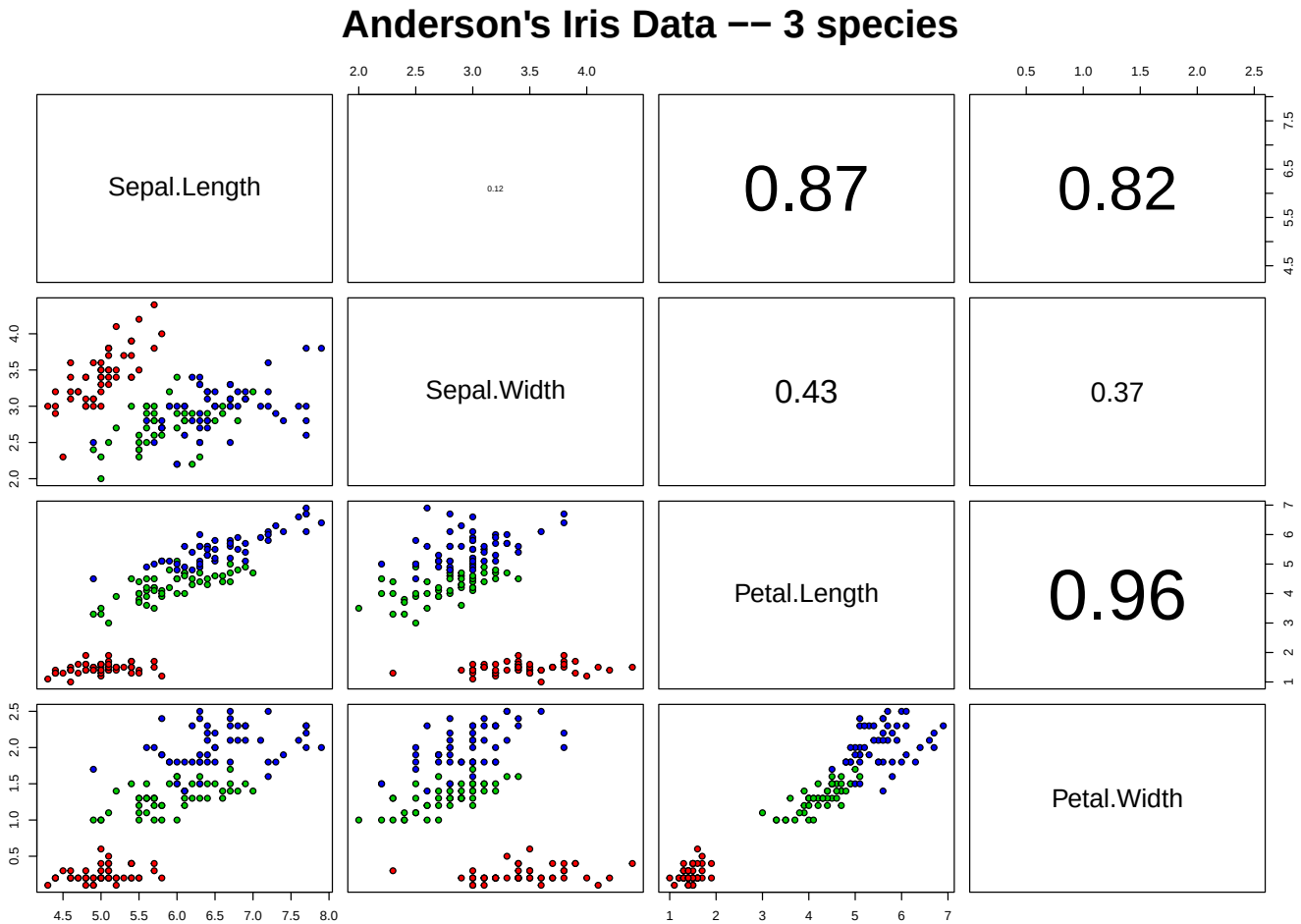
Quelques représentation pour décrire les Iris



Pour apprécier le lien entre Largeur et Longueur de Pétale

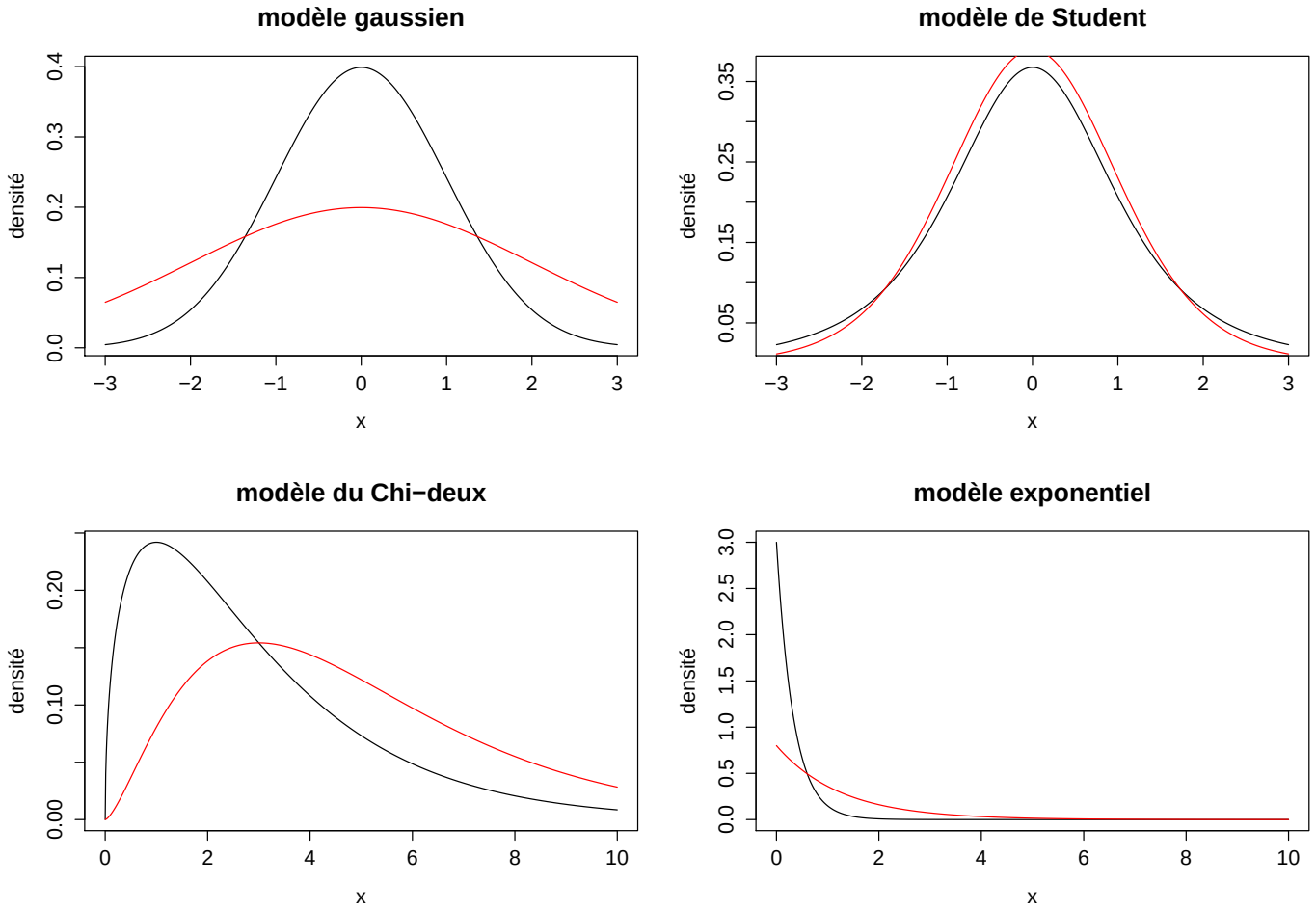


Et si on croise les 4 variables du jeu de données Iris avec spécification de l'espèce et des corrélations on obtient :

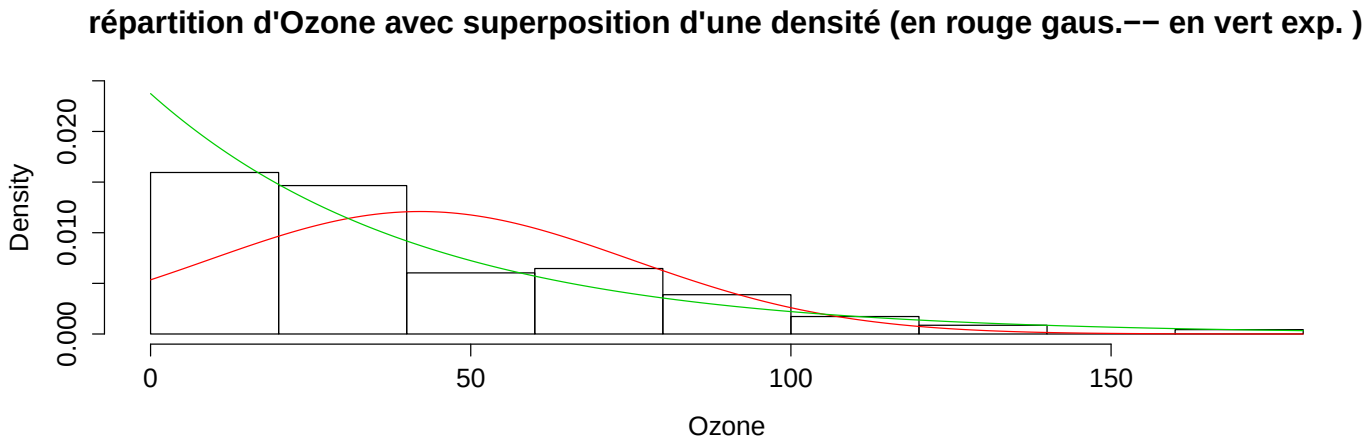


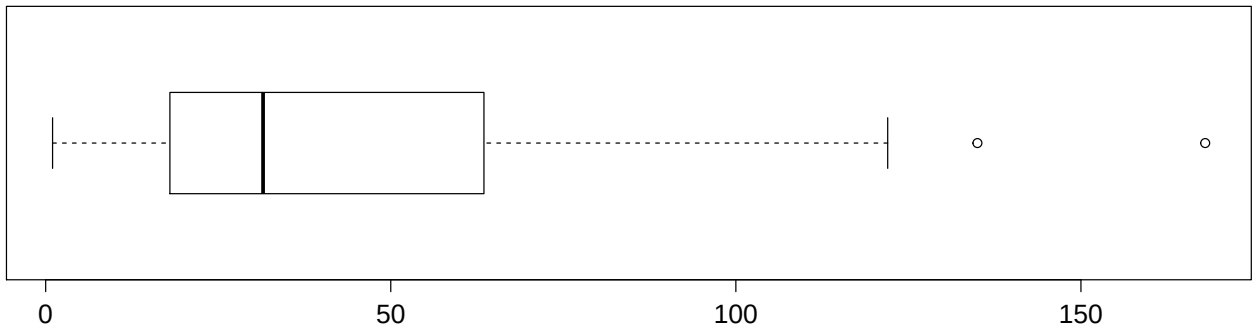
- $x_1, ..., x_n$: n réalisations indépendantes d'une variable X de densité $f_\theta(x)$
- proposer un modèle : choisir une famille de densité permettant d'approcher l'histogramme
- Trouver la famille de densités qui permettra d'ajuster au mieux la distribution observée (représentée par l'histogramme en densité). Quelques exemples de lois continues :
  - 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...

Densités des modèles continus usuels



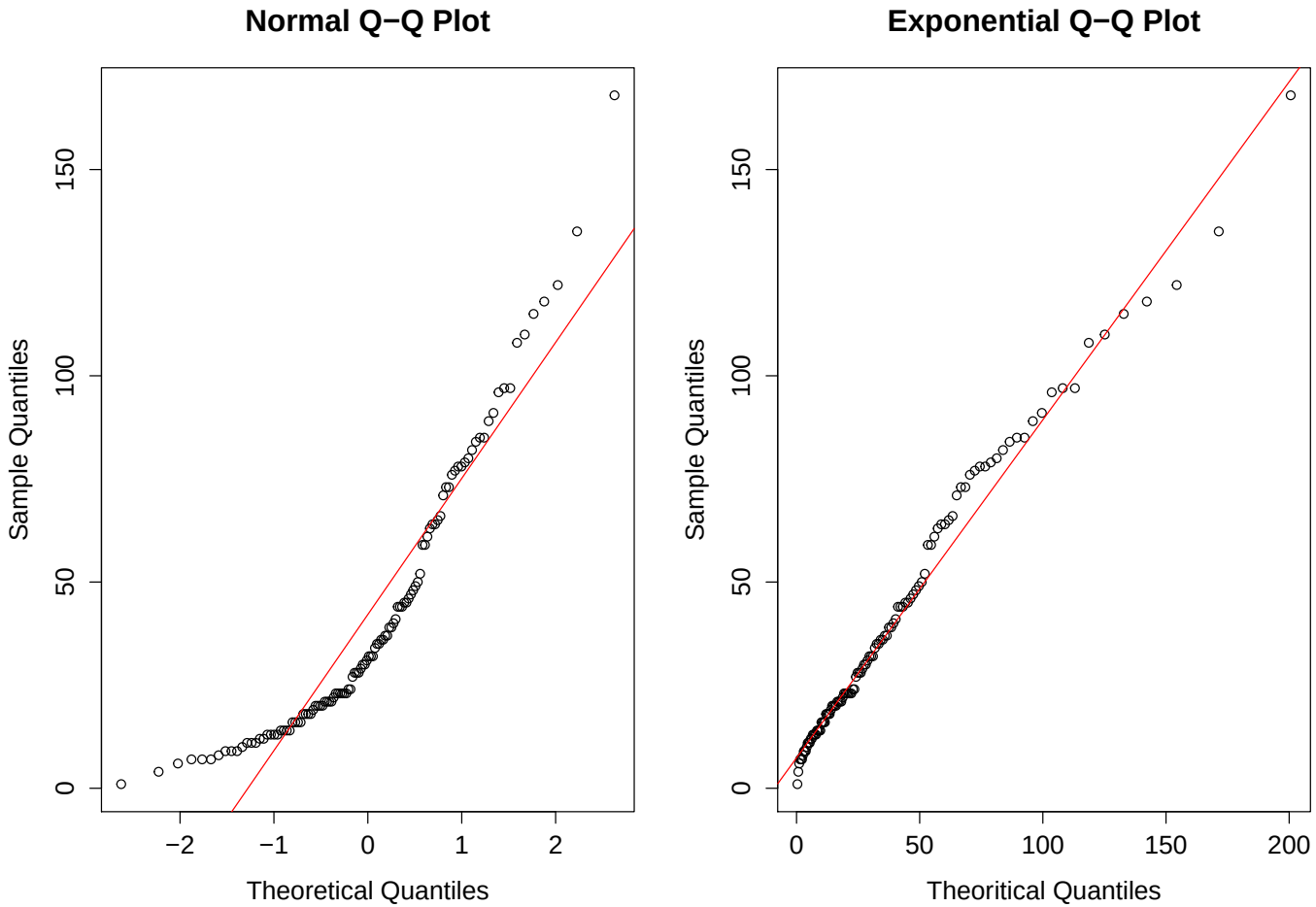
Quel modèle pour le taux d'Ozone des données `aiquality`?





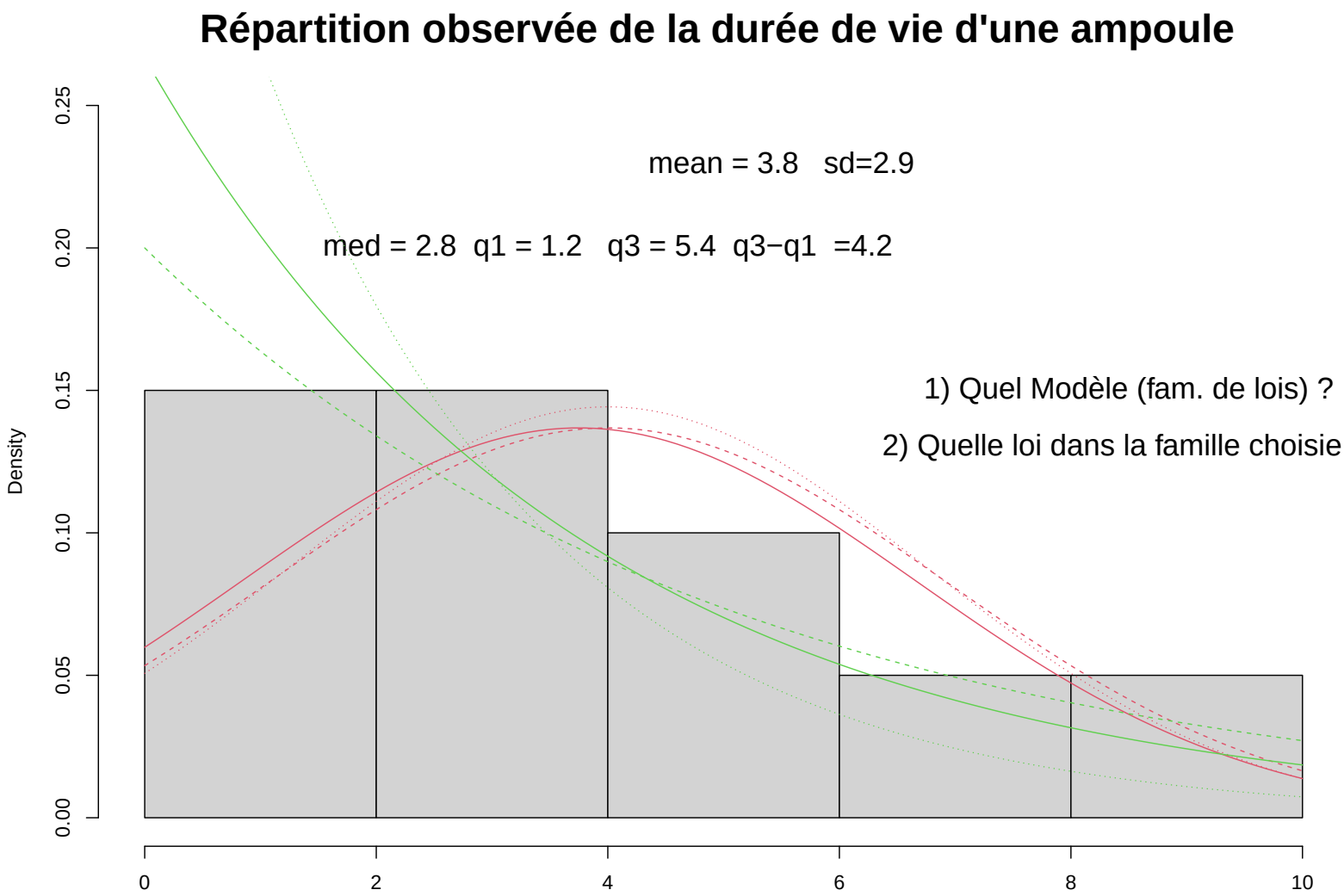
- le modèle gaussien est moins adapté que le modèle exponentiel.
- tracer le graphe appelé `qqplot()` : quantiles empiriques (ord.) vs quantiles théoriques du modèle cible (abs.) et y ajouter la droite de Henry (d'équation $y = \sigma x + \mu$ dans le cas gaussien). Si les points sont alignés autour de la droite on jugera le modèle cible proposé adapté (voir TP4).
- 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.



Les données de durées de vie de EX3

Quel modèle poser sur ces données ?



Utilisés dans ce cours deux outils (en **gras**) :

1. **Calculatrice programmable du bac** TI ou Casio (utilisée en CM et TD)
2. Logiciels statistiques : tableurs (stats de base), **logiciel R** (libre - outils graphiques très riches et méthodes statistiques nombreuses - utilisé en TP) , SAS (payant-pour de gros jeux de données), SPISS (payant - utilisé en sciences humaines), ...

Exercice à chercher pour le prochain cours

Soit G un générateur aléatoire de nombres choisis dans $\{1, 2, 3, 4\}$ avec des probabilités $\{p_1, ..., p_4\}$:

- tirage de deux séries de $n = 15$ réalisations pour un choix de la même loi de probabilité (inconnues pour vous mais spécifiées pour générer les données) fichier `de1.csv`.

$$x = (1, 4, 1, 3, 2, 1, 1, 4, 3, 1, 4, 1, 3, 3, 2)$$

$$y = (3, 3, 4, 2, 1, 3, 3, 2, 2, 3, 1, 1, 3, 3, 4)$$

- tirage d'une série de $n_z = 20$ réalisations pour un autre choix (ou pas ?) de la loi de probabilité. Fichier `de2.csv`.

$$z = (2, 3, 3, 1, 4, 4, 3, 1, 4, 4, 2, 4, 4, 2, 4, 4, 2, 3, 3, 3)$$

- de plus on note *de* la variable retournant "de1" si premier générateur et "de2" pour le second choix.

x, y, z sont quantitatives discrètes tandis que *de* est qualitative

avec la calculette Dans le menu Edit saisir

- x dans L1 // List1 : **(données brutes)**
- modalités de x dans L2 // List2 $m_k, k = 1, ..., 4$ **(données en effectifs)**
- effectifs associés : $n_k, k = 1, ..., 4$ **(données en effectifs)** dans L3 // List3
- On peut nommer chaque liste
- Avec le menu CALC/Stats 1-Var // CALC/1VAR afficher les résumés numériques de x avec deux méthodes : soit avec la première liste seule (données brutes), soit avec les deux listes suivantes (données en effectifs).

les résultats numériques

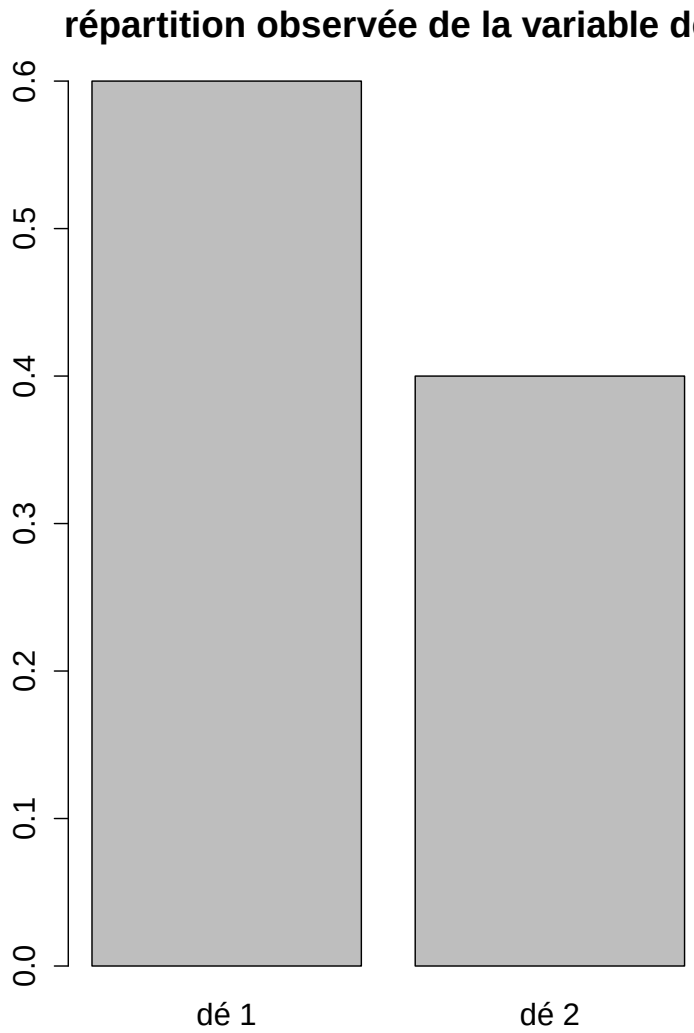
Remplir le tableau en effectifs et fréquences pour x :

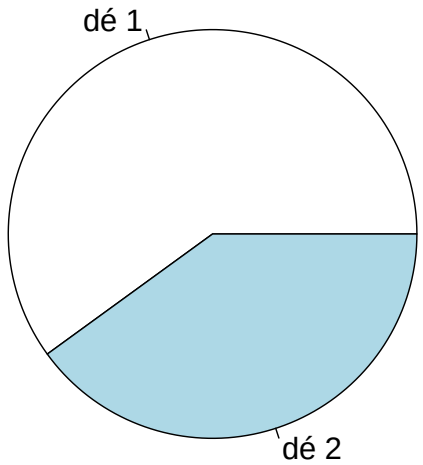
Modalités	1	2	3	4	total
Effectifs	6	2	4	3	15
fréquences %	40		26,67		
Fréquences cumulées %			80		

Et celui des résumés numériques pour x,y et z

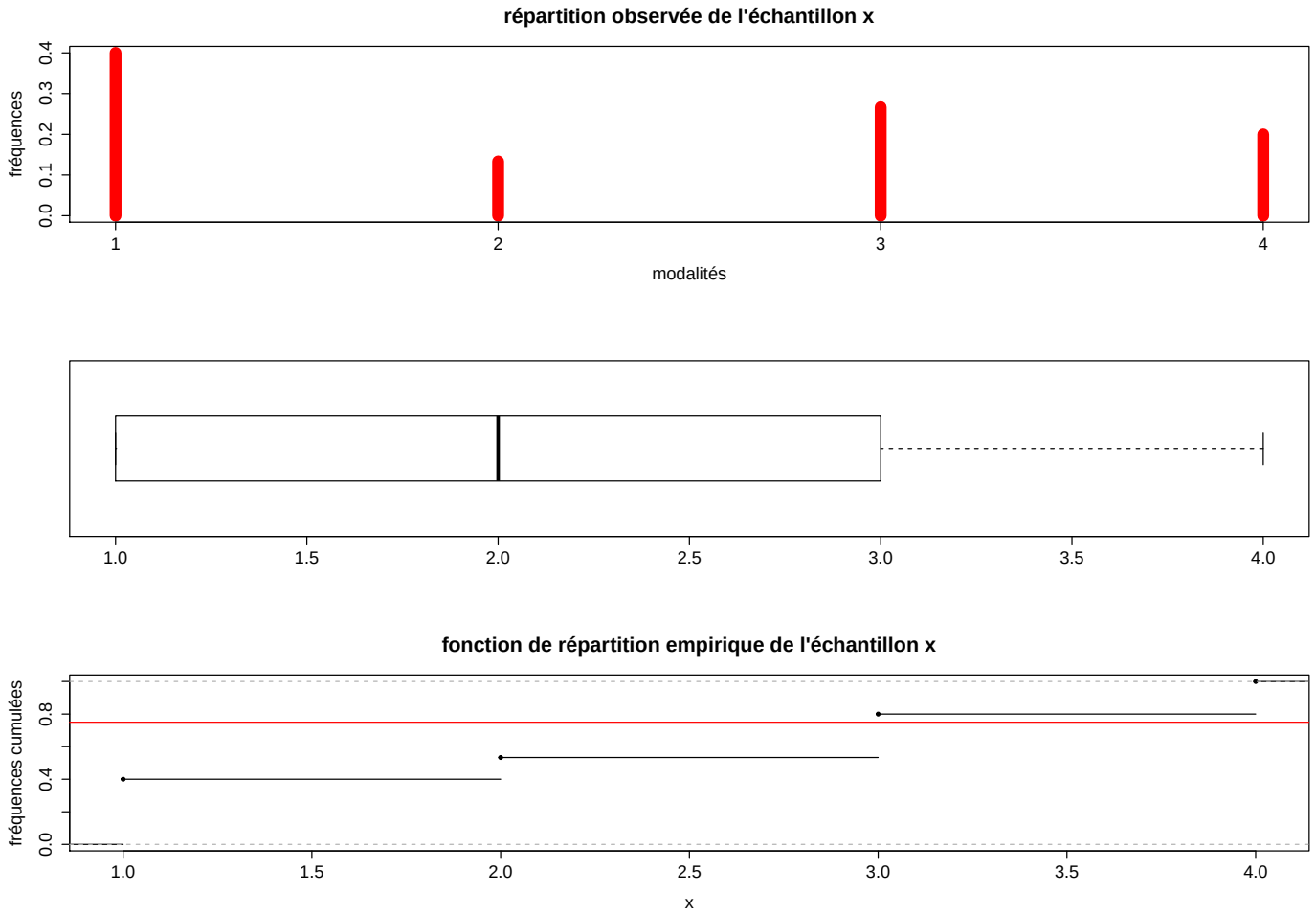
ech.	taille	moy.	méd.	q1	q3	v.e.	e.t.e.	min	max
x	15	2,27					1,18	1	
y						0,9156			
z			3	2	4				

les graphiques Variable *qualitative* (R : fonctions `barplot()` ou `pie()`)

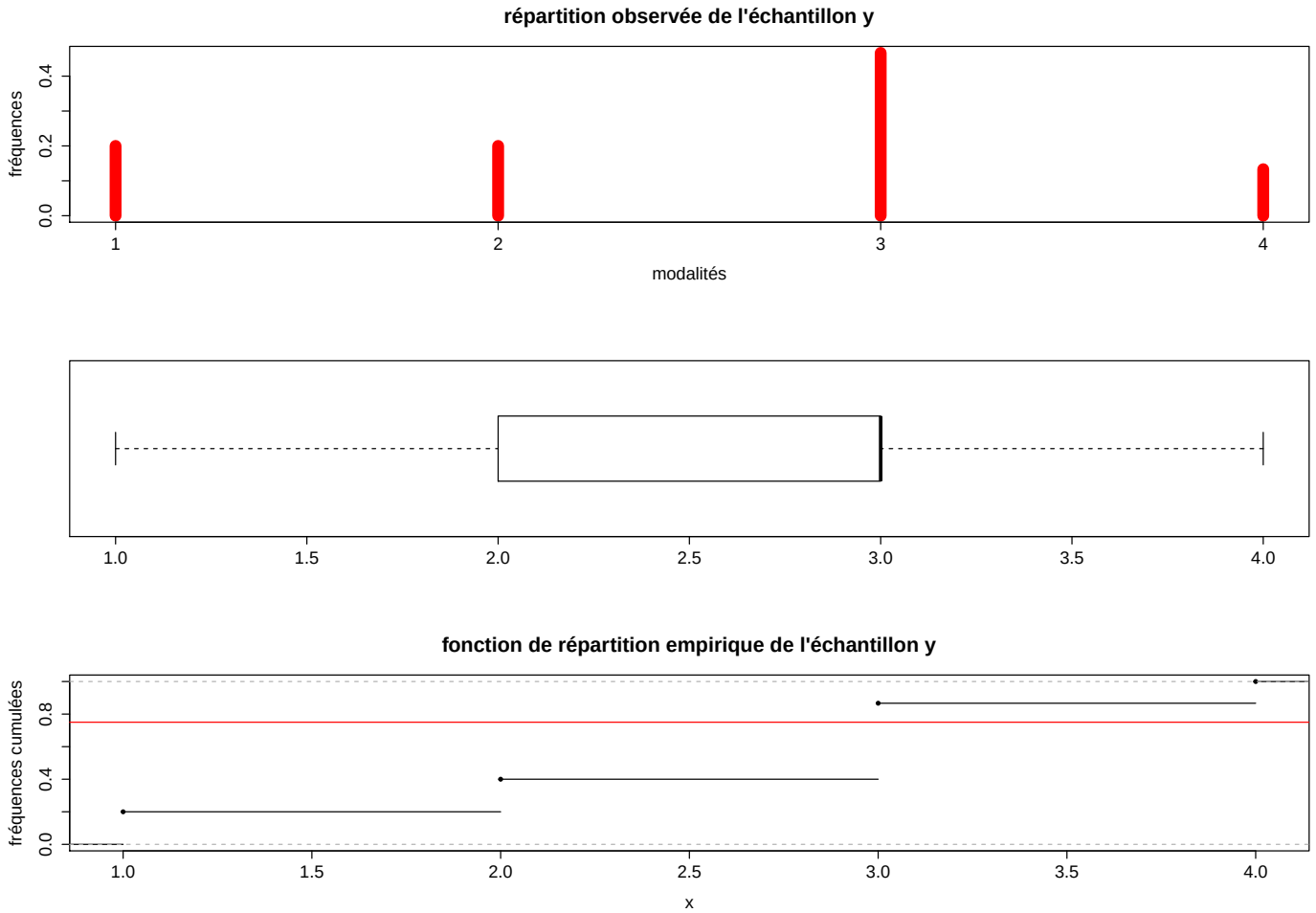




Variable *quantitative discrète* (R : fonctions `plot()` ou `boxplot()`)



Variable *quantitative discrète* (R : fonctions `plot()` ou `boxplot()`)



Variable *quantitative discrète* (R : fonctions `plot()` ou `boxplot()`)